EUROPEAN LANGUAGE GRID

D5.1

Identification and collection of existing datasets, models, identified gaps and plans (version 1)

Authors:	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Cristian Martinez (ELDA), Penny Labropoulou, Miltos Deligiannis, Stelios Piperidis (ILSP)

Dissemination Level:	Public
Date:	30-04-2020

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 36 months
Deliverable number	D5.1
Deliverable title	Identification and collection of existing datasets, models, identified gaps and plans (version 1)
Type	Other (Report + Data Sets)
Number of pages	44
Status and version	Final – Version 1.0
Dissemination level	Public
Date of delivery	Contractual: 30-04-2020 – Actual: 30-04-2020
WP number and title	WP5: Grid Content – Language Resources, Datasets, and Models
Task number and title	Task 5.1: Identification and collection of existing datasets and resources to make them available through the ELG
Authors	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Cristian Martinez (ELDA), Penny Labropoulou, Miltos Deligiannis, Stelios Piperidis (ILSP)
Reviewers	Georg Rehm and Katrin Marheinecke (DFKI), Gerhard Backfried (SAIL)
Consortium	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Institute for Language and Speech Processing (ILSP), Greece
	University of Sheffield (USFD), United Kingdom
	Charles University (CUNI), Czech Republic
	Evaluations and Language Resources Distribution Agency (ELDA), France
	Tilde SIA (TILDE), Latvia
	Sail Labs Technology GmbH (SAIL), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Alexandru Ceausu
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Dr. Georg Rehm, DFKI GmbH georg.rehm@dfki.de Phone: +49 (0)30 23895-1833 Fax: +49 (0)30 23895-1810 http://european-language-grid.eu © 2020 ELG Consortium

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CC	Creative Commons
CMDI	Component Metadata Infrastructure
DCAT	Data Catalog Vocabulary
DOIP	Diagnostics over Internet Protocol
ELG	European Language Grid
FBK	Fondazione Bruno Kessler
ICT-29b)	Six Horizon 2020 projects related to Language Technology
ILC	Institute for Computational Linguistics “A. Zampolli”
ILSP	Institute for Language and Speech Processing
IPIPAN	Institute of Computer Science, Polish Academy of Sciences
ISLRN	International Standard Language Resource Number
LR or LRs	Language Resources or Language Resource
LT or LTs	Language Technology or Technologies
NCCs	National Competence Centres of ELG
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCR	Optical Character Recognition
OLAC	Open Language Archives Community
OMTD-SHARE	OpenMinTeD project
PID	Persistent Identifier
R1a, R1	Pre-release (R1a), and Release 1 of ELG
SPDX	Software Package Data Exchange
TMX	Translation Memory Exchange
WMT	Workshop on Machine Translation
XML	Extensible Markup Language
XSD	XML Schema Definition
XSLT	Extensible Stylesheet Language Transformations

Abstract

Deliverable D5.1, the report on "Identification and collection of existing datasets, models, identified gaps and plans", describes the work carried out to populate the European Language Grid (ELG) platform with Language Resources (LRs). After a brief introduction recalling the objectives, the document elaborates on the approaches and methodologies to identify repositories of LRs and Language Tools. The repositories that were selected for ingestion in the ELG platform are described in detail via their metadata elements. The ingestion process required a careful and thorough conversion from the various metadata schemas into the one designed and implemented for ELG. In this first set of resources to ingest, we considered three important sources: The ELRA catalogue, the ELRC-SHARE repository, and several nodes of the META-SHARE network of repositories. These repositories exploit metadata elements based on the ones developed in the framework of the META-NET project that built the META-SHARE network including its catalogues. The conversions are now operational and can be customised to ingest new repositories for upcoming releases of the ELG platform. The report also lists the resources that are made available within this first release of the ELG platform.

1 Introduction

D5.1 reports on the work carried out towards the identification and ingestion of datasets and language models for ELG's pre-release (Release 1 alpha, R1a) and first release (Release 1, R1). This work consisted in the identification of sources (inventories and repositories), language resources and models, their analysis, selection of elements to be ingested, as well as in the conversion of their metadata descriptions and ingestion.

The approach and methodology followed in the identification stage are also described, as well as the different steps until ingestion into ELG. Section 2 describes the approach and methodology followed for the identification. Section 3 lists the repositories identified. Section 4 focuses on the data identified, which datasets were pre-selected, and which have been ingested into R1. Section 5 details the metadata conversion work done for the different repositories targeted for R1. Section 6 elaborates on the gaps identified during the whole process and the experience acquired during this phase, and Section 7 provides some insights on the next steps to take. Several annexes give more details about the resources and the conversion processes.

The term "Language resource" (LR, LRs) is used for resources composed of linguistic material deployed in the construction, improvement or evaluation of Language Technologies (LT, LTs), but also, in a broader sense, in language and language-mediated research studies and applications; examples include datasets of various types, such as textual, multimodal or multimedia corpora, lexical data, grammars, language models, etc. The term is often used in the bibliography and related initiatives with a broader meaning, encompassing also the (a) tools and services used for the processing and management of datasets, and (b) standards, guidelines and similar documents that support the research, development and evaluation of LTs. In the ELG metadata model, we use the term as first defined for the META-SHARE metadata model, i.e., including both data resources and LT tools/services. The alternative term "Language Resource/Technology" is also used in the context of ELG.

2 Identification Approach and Methodology

The European Language Grid (ELG) aims to become the primary marketplace for the European LT landscape. The organisations range from commercial to non-commercial, include research centres and companies, as well

as initiatives and infrastructures, among others. Linking all these players and supporting them in their interaction is a two-fold mission, which involves (1) helping them make their tools, services and data available and (2) establishing the means for them to find and have access to those they may require in their work. In order to receive a comprehensive set of existing repositories, ELG not only relied on resources provided by members of the ELG consortium but also reached out to its National Competence Centres (NCCs).

2.1 Tasks Carried out by the Consortium

In the first step, ELG completed the first stage of Task 5.1 "Identification and collection of existing datasets and resources to make them available through the ELG" by examining existing and available inventories and repositories of all potential LT/LR providers and users. This identification work has been part of a collaborative task within the ELG consortium, counting on contributions from the different partners. ELDA completed this first identification wave (see Section 3). The current results will be completed during the next stages of the project. The description and identification of inventories and repositories comprises the elements shown in Table 1.

Minimal set	Metadata element
Definition of a minimal set of metadata elements for the description of repositories:	• Repository Name • URL • Partner who identified the repository, • Repository description
Definition of an optional set of metadata elements for the description of repositories (if information available):	• Languages • Typology of LRs (speech, written, multimodal, mixed) • Possible applications • Project(s) in which the repository was established and/or used • Domain(s) • Legal information • Restrictions of use (research, commercial, etc.) • Repository owner's name • Type of owner (academic, industry, other) • Repository owner's physical address • Repository owner's mailing address • Repository owner's phone number • Repository owner's email address • Repository owner's country, comments
Identification of the major LR&LT Inventories	• META-SHARE • ELRA • LREMAP • CLARIN • ELRC-SHARE • Open Data Portals (EU, National) • ICT-29b) projects
Identification of the major portals outside of the EU	• USA, e.g., LDC[1] • South Africa, e.g., SADiLaR[2] • China, e.g., LBDA[3] • etc.
Identification of the major national programs	• Spanish language technology program • Danish and Norwegian actions • Involvement of the ELG's NCCs

Table 1: Definition of sets of metadata elements

[1] Linguistic Data Consortium (LDC, https://www.ldc.upenn.edu)
[2] The South African Centre for Digital Language Resources (SADiLaR, http://www.sadilar.org)
[3] http://www.lbdalliance.com/yydsjlmen/index.html

Regarding the typology of LRs searched for, all types and modalities which are useful for some sort of LT application have been considered. These features comprised corpora, lexicons, terminologies, and derived resources (such as Language Models for Automatic Speech recognition or Translation Memory eXchange models for Machine Translation), and also focused on media such as:

- Speech/Audio
- Text
- Video/Audio-visual
- Images, OCR
- Sign language datasets (images, videos)

2.2 Tasks Carried out by the ELG National Competence Centres

In addition to the work done by the members of the consortium, DFKI, as the partner responsible for the NCCs, carried out a survey to gather more input from the NCCs and from the partners, often related to their local and regional repositories. Some of the reported repositories were already identified by the consortium (mostly META-SHARE and CLARIN nodes) but new ones have been listed and, moreover, with very extensive documentation provided by the NCCs (content, contacts, etc.). The new ones are described here and could be ingested in ELG in one of the coming releases. These are the following ones.

- OPUS – the open parallel corpus (http://opus.nlpl.eu): this is one of the largest and growing collections of translated texts from the web. The objective of the OPUS project is to crawl, convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package.
- Corola (http://corola.racai.ro): RACAI/ICIACOROLA is not a repository per se but rather a large corpus of Romanian text and speech data, distributed in packages and accompanied by CMDI-compliant metadata. The texts are classified according to general domains and are pre-processed (morpho-lexically tagged, lemmatized). The speech part contains almost one million words (over 100 hours of recordings). Speech data are transcribed and aligned.
- RELATE (http://relate.racai.ro): This is a Portal of Romanian Language Technologies but comprises sections with useful assets to ELG (both LRs and LTs). It comprises a collection of tools and resources for dealing with the Romanian language that may be beneficial to ELG users at a later release of the ELG platform.
- PARLI portal (http://parli.di.unito.it/resources_en.html): This is also a portal with a section that lists many resources developed by the Italian research community for the Italian language and that may benefit from the ELG services but also bring more assets. It consists of mostly textual data.
- The Department of Computational Linguistics (DCL) of the Institute for Bulgarian Language resources (https://dcl.bas.bg/en/resursi/): This is another very important repository of downloadable language resources for Bulgarian. It is important as Bulgarian may be considered as a low-resourced language and hence could greatly benefit from ELG services. This repository comprises corpora, tools and services, mostly available under Creative Commons licences (https://creativecommons.org, CC) or in the public domain.

This contribution from the NCCs is highly valuable. We will continue to work with them in this respect very closely once the ELG platform is fully operational.

3 Identified Repositories

3.1 Repositories for the European Language Grid

As mentioned in Section 2, the ELG partners have contributed to the identification and compilation of a list of resource repositories, most of which will be also part of the data provision for ELG. The list of repositories can be seen in Annex A. Only the first two minimal metadata fields (Name and URL) are provided, which are the basis of the classification. A preliminary description has been added to illustrate the ongoing analysis work that is taking place and that will be finalised for R2.

This list comprises different types of portals, such as those storing data from evaluation campaigns or shared tasks (e.g., WMT resources)[4], large catalogues of language resources (e.g., ELRA), networks of LR repositories (e.g., META-SHARE nodes), data banks, initiatives collecting or supporting the collection of language data, etc.

These repositories are going to be the basis for language resource provision in ELG. An initial selection has taken place for the first releases of the platform (R1a and R1) in order to provide a set of resources for its first users (in particular, for the pilot projects at this stage), but the remaining repositories are being analysed and prioritised for integration in R2, taking into account the following types of information:

- Relevance of their content for ELG and its services.
- Usability information (conditions of use/sharing, licensing, etc.).
- LR typology (trying to cover different modalities as well as filling up detected gaps).
- Metadata schema used (closer conversions for interesting and shareable content may gain priority).
- Languages covered (as with modalities, trying to cover targeted languages from the three defined groups[5] and trying to reach a balance in coverage and gap filling.
- Number of datasets or models contained within the repository (if large numbers of relevant and easier-to-obtain datasets are available, this will be a priority criterion).

3.2 Repository Priorities for ELG Release 1 alpha and Release 1

Following the identification work described above, three repositories have been selected for ingestion in the first two releases R1a and R1. These repositories are run by ELG consortium partners: ELRA[6], ELRC-SHARE[7] and three nodes from META-SHARE (those managed by DFKI[8], ELDA[9] and ILSP[10]). The reason behind this choice has been strategical, as a proof of concept for both resource availability and metadata conversion, given that partners are more familiar with the content and metadata schema within these repositories.

Regarding META-SHARE, only those nodes run by Consortium partners have been considered until collaboration agreements are reached with those responsible for the remaining nodes. These agreements are of a rather

[4] http://www.statmt.org/wmt20/ (Workshop on Machine Translation)
[5] Following ELG's specifications, languages have been classed into three groups: A – Official EU Languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish. B – Other EU languages; languages from EU candidate countries and Free Trade Partners: Albanian, Basque, Catalan, Galician, Icelandic, Norwegian, Scottish Gaelic, Welsh, Serbian, Turkish, Ukrainian. C– Languages spoken by EU immigrants; languages of important trade and political partners: Afrikaans, Arabic, Berber, Cebuano, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Kurdish, Latin, Malay, Pashto, Persian (Farsi), Russian, Tamil, Vietnamese.
[6] http://catalogue.elra.info
[7] https://elrc-share.eu
[8] http://metashare.dfki.de
[9] http://metashare.elda.org
[10] http://metashare.ilsp.gr:8080

flexible nature where leaders can choose an optimal integration-to-ELG approach which can imply either migrating the content of their nodes or maintaining the node but making its content accessible through ELG. Discussions have already started with those heading the META-SHARE managing nodes with a very positive response to integrate into ELG. The afore-mentioned 2-case scenarios have been suggested and ELG is going to start working on the integration of these nodes (FBK[11], IPIPAN[12] and ILC[13]) as soon as the R1 release has been successfully launched at the beginning of May 2020. Both, these nodes and the rest of the META-SHARE nodes are part of the target data provisions for R2.

Furthermore, it is envisaged that LINDAT[14] is part of the repositories following ELRA, ELRC-SHARE and META-SHARE nodes, with its conversion/harvesting to be accomplished for R2. Work has already started on the analysis of its metadata elements and their mapping to ELG elements. Once R1 is concluded, this mapping will be continued in close collaboration with LINDAT's managers, to identify metadata mismatches and harvesting, as well as data integration.

The Open Language Archives Community repository (OLAC)[15] is also under analysis for R2. OLAC provides a compilation of 59 source inventories and repositories which are converted to its metadata and then harvested. This catalogue is very rich in terms of finding other sources, but as it aggregates from other repositories, it will be used only to identify repositories of interest to ELG (and hence will not be harvested). OLAC will be very useful for this purpose as it describes resources with a simplified metadata schema that help decide on the importance of the repositories for ELG. In this case, deeper descriptions are required (as needed by ELG), and we will need to refer to the original records.

3.3 The ELG Metadata Model

The ELG metadata model (also referred to as ELG-SHARE) is used for the description of all entities of interest to the ELG target users (for a detailed description, see D2.4)[16]. It constitutes the backbone of the ELG catalogue, which brings together language processing services and tools, LRs (datasets of different types and media) as well as actors and activities, related to LT (see Figure 1).

The model caters for the description of the ELG core entities, i.e.,

- LT tools/services, covering all software that performs language processing and/or any LT-related operation (e.g., basic processing tools, applications, web services etc. that perform annotation, Machine Translation systems, Automatic Speech Recognition and Synthesis, Information Retrieval, etc.).
- *Corpora* (also: datasets), defined for our purposes as structured collections of pieces of data (textual, audio, video, multimodal/multimedia, etc.), typically of considerable size and selected according to criteria external to the data (e.g., size, type of language, type of text producers or expected audience, etc.) to represent as comprehensively as possible the object of study.
- *Lexical/conceptual resources*, i.e., resources (such as terminological glossaries, word lists, semantic lexicons, ontologies, gazetteers, etc.) organized on the basis of lexical or conceptual units (lexical items,

[11] The Fondazione Bruno Kessler (FBK), Trento, Italy runs the http://metashare.fbk.eu node.
[12] The Institute of Computer Science Polish Academy of Sciences runs the http://metashare.nlp.ipipan.waw.pl node.
[13] The Institute for Computational Linguistics «A. Zampolli» of the Italian CNR runs the http://metashare.ilc.cnr.it node.
[14] LINDAT is the CLARIN Centre for Language Research Infrastructure in the Czech Republic.
[15] OLAC, the Open Language Archives Community, an international partnership for creating a worldwide virtual library of language resources, http://www.language-archives.org
[16] Penny Labropoulou, Katerina Gkirtzou, Miltos Deligiannis, Dimitris Galanis, Maria Gavriilidou, Stelios Piperidis, Georg Rehm, Maria Moritz, Andrés Garcia Silva: ELG deliverable D2.3: Metadata schema (August 2019)

terms, concepts, phrases, etc.) with their supplementary information (e.g., grammatical, semantic, statistical information, etc.).

- *Language descriptions*, i.e., resources aiming to describe a language or some aspect(s) of a language via a systematic documentation of linguistic structures (e.g., computational grammars, statistical and machine learning-computed language models).

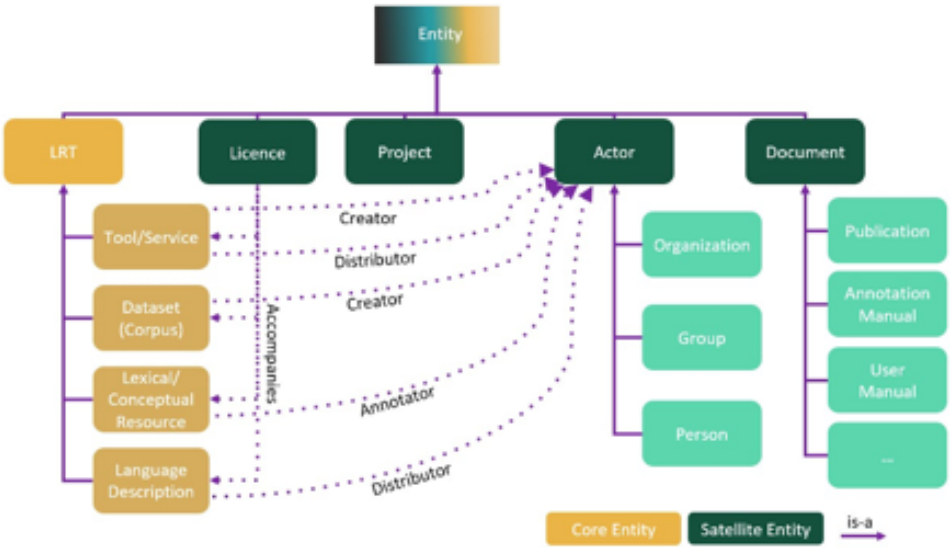

Figure 1: ELG entities in the metadata model

It also covers entities involved in the production and usage of LTs/LRs and, in general, LT activities, namely *actors* (organizations, groups and persons), *documents* (e. g., user manuals, publications, etc.), *projects* and *licenses/terms of use*.

WP5 focuses on the collection and integration into the ELG of data resources only, i.e., those corresponding to the subclasses of corpora, lexical/conceptual resources, and language descriptions. However, their description entails the description of related resources and entities. Therefore, we include here a brief overview of the whole model and its main features.

The ELG model includes a large number of metadata elements grouped along three key concepts: resource type, media type and distribution. The resource type element distinguishes Language Resources and Technologies in the four classes presented above. Media type refers to the form/physical medium of a data resource (or of its parts, in the case of multimodal resources), i.e., text, audio, image, video and numerical text (used for biometrical, geospatial and other numerical data). Finally, distribution, following the W3C DCAT[17] vocabulary, refers to the physical form of the resource that can be distributed and deployed by consumers. For instance, software resources may be distributed as web services, executable files or source code files, while data resources as PDF, CSV or plain text files or through a user interface. Administrative and descriptive metadata are mostly common to all LRs and LTs, while technical metadata differ across resource and media types as well as distributions. Figure 2 provides an example with part of the metadata model for tools/services.

[17] https://www.w3.org/TR/vocab-dcat-2/

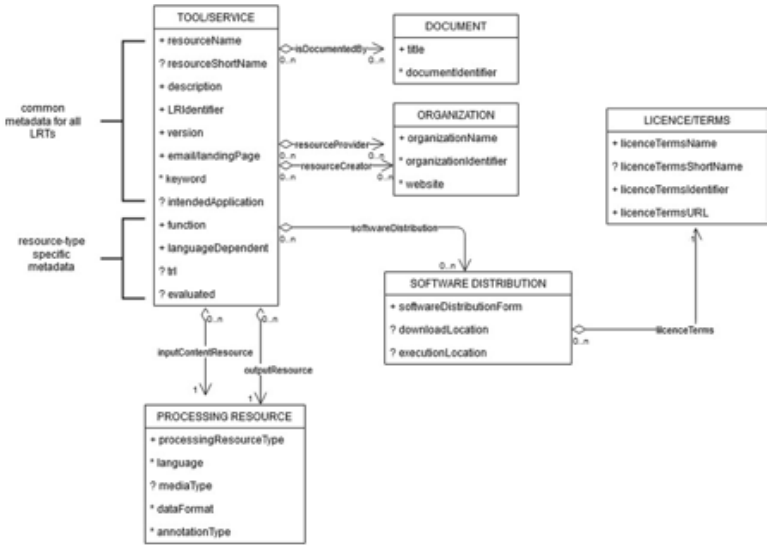

Figure 2: ELG metadata model for Tools/Services

The design and implementation of the model (except for the billing module) have now been completed. Nevertheless, we anticipate updates and improvements for the next releases of the platform, taking into account user feedback and technical requirements of the platform as its implementation progresses. The model is implemented in the form of an XML Schema Definition (XSD), with its elements linked to entities from two ontologies, namely the META-SHARE ontology[18], which includes the majority of elements and controlled vocabularies, and the OMTD-SHARE[19] one, reserved for the controlled vocabularies of LT categories (also referred to as "LT taxonomy"), data formats and methods.

Given that the ELG model aims to cover the description of the whole lifecycle of LRs and LTs, it includes an extensive set of elements, which makes the process of creating metadata records effort-consuming. To ensure flexibility and uptake, metadata elements are distinguished into *mandatory, recommended* and *optional* ones. A minimal version with only the mandatory and recommended elements has been created based on a careful selection of elements deemed necessary or important for the EL G purposes.

Finally, an important feature of the ELG model that has been instrumental in the prioritisation of repositories for R1 is the fact that it draws upon and extends the META-SHARE schema and its profiles (Figure 3).

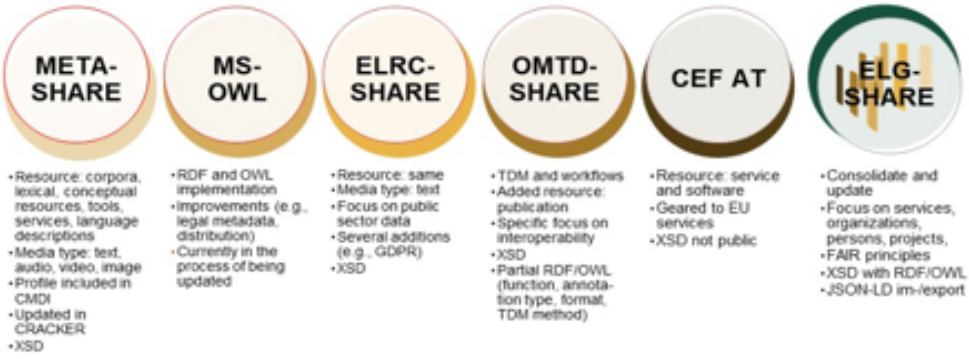

Figure 3: Evolution of the ELG metadata schema

[18] https://w3id.org/meta-share/meta-share/
[19] The OMTD-SHARE ontology for resources (https://w3id.org/meta-share/omtd-share/) is related to the Text and Data Mining initiated in the framework of the OpenMinTeD project (https://www.openminted.eu)

4 Identified Language Resources

For R1a (April 2020) about 200 LRs per repository (ELRA, ELRC-SHARE and META-SHARE) were pre-selected. Their licensing and distribution conditions were analysed and the details for those pre-selected as being free and legally uploadable are described in the subsections below. The remaining LRs that need to be further cleared with either their providers or node managers (for instance, in META-SHARE's case with regard to its META-SHARE proprietary licenses) will be the objective of R2. However, out of these almost 600 pre-selected resources, a prioritisation has been done based on what could be handled by the ELG platform both legally and technically for both R1a and R1. This is further explained in Section 4.2.

4.1 Selected Language Resources from Priority Repositories

R1 is focusing on the improvement of R1a by solving issues encountered and making procedures more agile and robust. This is why no further resources have been added. This will also be detailed in Section 4.2. Regarding the technical prioritisation approach, LRs have been classified per resource type and different types have been considered for conversion and uploaded.

Given the heterogeneous nature of the metadata models used by the different repositories, the conversion mechanisms to the ELG model were carried out and they allowed to identify and correct a number of metadata issues within our models. The first resource type addressed was "Text", followed by "Audio". The type "Evaluation package" was left for a later phase of the project given their nature and usage. Evaluation packages are made up of different sets of data used in evaluation campaigns and challenges. As the outcome of an evaluation campaign or a shared task, organizers compile into one single package the LRs (used for training, for development and for testing), with the evaluation protocols (e.g., definition of the task to achieve), the scoring tools, the results of the evaluation campaign, etc. All these items have to be described as a single entry in our catalogues. The ELG partners are discussing the best way to handle this category that links data, tools, scoring metrics, etc. For R1, evaluation packages are not considered.

The following sections provide an overview of what has been pre-selected and prepared for ingestion from the ELRA, ELRC-SHARE and META-SHARE node repositories. Besides, these sections show the filtering that has been done to these pre-selections and the reasons for such decisions.

4.1.1 ELRA

The first batch of pre-selected LRs from ELRA consisted of 103 datasets that are free for use, and that can be split into the LR types listed in Table 2.

Media Type	Resource Type	#LRs
Text	Corpus	49
	Lexical conceptual database	6
Audio	Corpus	48
Video	Corpus (combined with text)	1
	Corpus (combined with audio)	2

Table 2: ELRA – LRs pre-selected for ingestion into ELG

The terms of use of these datasets refer to the nature of the institution to use the data (academic, commercial). These different conditions will help us understand the sharing restrictions on the one hand, and the needs

to expand metadata and functionalities within ELG on the other hand. The latter are foreseen for the R2 release planned at the end of 2020. 22 out of these 103 LRs have been ingested for R1 due to the mentioned technical work to be done. The list of currently ingested ELRA LRs for R1 can be seen in Section 4.2.

4.1.2 ELRC-SHARE

This repository provides 187 LRs that can be used for free. These LRs can be split among the resource types and linguality types shown in Table 3. All ELRC-SHARE LRs are textual datasets.

Resource Type	Linguality Type	#LRs
Corpus	Monolingual	2
	Bilingual	168
	Multilingual	10
Lexical conceptual database	Bilingual	6
	Multilingual	1

Table 3: ELRC-SHARE – LRs pre-selected for ingestion into ELG

The licensing conditions for these 187 datasets allow for their ingestion in an immediate manner, both from the legal and technical points of view. In addition, all the corresponding conditions and restrictions can be described with the current metadata schema used within ELG. The licensing conditions are as follows:

Licenses	#LRs
CC-BY-2.0	1
CC-BY-2.5-SE	1
CC-BY-3.0	4
CC-BY-4.0	30
CC-BY-NC-4.0	4
CC-BY-ND-3.0	1
CC-BY-SA-3.0	5
CC-BY-SA-4.0	6
openUnder-PSI	90
publicDomain	45
TOTAL	187

Table 4: ELRC-SHARE – LRs available under open licenses

4.1.3 META-SHARE

Regarding the META-SHARE nodes maintained by DFKI[20] and ILSP[21], which were selected for R1, 212 datasets have been initially identified as suitable for ELG, belonging to the resource types detailed below:

Node	Resource Type	#LRs
DFKI	Corpus	7
ILSP	Corpus	182
	Linguistic description	5
	Lexical conceptual database	18

Table 5: META-SHARE – LRs pre-selected for ingestion into ELG

[20] http://metashare.dfki.de
[21] http://metashare.ilsp.gr:8080

The first analysis showed that 71 out of those 212 offer open sharing conditions that allow ELG and the META-SHARE partners to agree on an immediate integration. This is the subset of those initially selected that are provided to ELG for R1. Table 5 lists the pre-selected LRs for the prioritised META-SHARE nodes but does not contain ELDA's node as these resources are provided directly through the ELRA Catalogue.

4.2 Language Resources Provided for ELG Release 1 alpha and Release 1

As seen in Section 4.1, three series of datasets were originally pre-selected from the three prioritised repositories to be converted and ingested into the ELG platform for R1a. However, only ELRC-SHARE allowed for the insertion of the full selected list given that its resources met the following conditions: a) their licensing conditions allowed it (all data were shared under CC-BY licenses, they were open under PSI, or they were public domain data), and b) their metadata elements were compatible and fully covered by the ELG metadata schema.

ELG offers different ways to collaborate for interested institutions, allowing them to choose whether they wish to integrate their data and metadata descriptions in the platform or simply the metadata descriptions. ELRC-SHARE follows the second option, provided that this repository is already hosted by ILSP, on behalf of the EC, and datasets should remain there for at least the duration of the ELRC contracts. For that reason, the master copies of the 187 LRs provided to ELG remain within ELRC-SHARE and their metadata records are available through the ELG platform, allowing to be redirected for download to the ELRC-SHARE catalogue pages.

The list of imported datasets from ELRC-SHARE to ELG can be seen in Table 6. To ensure that each resource on the ELRC-SHARE repository gets a PID, the ELRC consortium has assigned each resource an International Standard Language Resource Number (ISLRN[22]) that is listed here, together with the languages of the resource.

Resource name	ISLRN	Languages
English-Estonian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	367-945-013-309-2	et–en
English-Lithuanian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	133-724-111-130-7	en–lt
English-Danish EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	034-297-263-067-2	en–da
English-Latvian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	704-517-283-753-9	lv–en
SIP Internal dictionary (Processed)	495-192-259-373-0	de–en–fr
Czech Banking Association Terminology (Processed)	620-065-803-223-7	en–cs
Czech Association of Medical Physicists – Physics Glossary (Processed)	669-761-211-618-8	en–cs
English-Swedish parallel corpus from the translation of 'Sweden a Pocket Guide' book (Processed)	790-580-207-032-9	en–sv
Bilingual hr-en parallel corpus from Croatian Mine Action website (Processed)	789-620-257-493-3	en–hr
Bilingual collection of documents about the Cyprus Problem (Processed)	391-837-431-539-1	en–el
Bilingual documents Bulgarian-English in the field of ICT and Transport (Processed)	942-857-416-126-2	en–bg
Bilingual documents Bulgarian-English in the field of open data, broadband and information society (Processed)	812-088-133-045-4	en–bg
Bilingual hr-en parallel corpus from the National and University Library in Zagreb website (Processed)	196-404-604-094-3	en–hr
English-Slovak corpus of annual reports on immigration and asylum policies from the EMN National Contact Point for the Slovak Republic website (Processed)	496-687-732-206-0	en–sk

[22] http://www.elra.info/en/islrn/overview/

English-Slovak corpus of annual reports from the Slovak National Centre for Human Rights website (Processed)	984-293-484-233-4	en–sk
EUIPO – IP case law French-English (Processed)	677-401-941-686-1	en–fr
EUIPO – IP case law Spanish-English (Processed)	268-462-170-942-3	en–es
EUIPO – IP case law German-English (Processed)	510-915-622-048-3	de–en
EUIPO – IP case law Italian-English (Processed)	896-949-472-998-2	en–it
Croatian-English corpus with Acts on Biological and Landscape Diversity and Environmental Protection (Processed)	095-764-087-898-1	en–hr
EUIPO – list of goods and services German and English (Processed)	522-601-732-320-1	de–en
EUIPO – list of goods and services German and Spanish (Processed)	879-151-530-310-4	de–es
EUIPO – list of goods and services German and French (Processed)	372-893-655-274-0	de–fr
EUIPO – list of goods and services German and Italian (Processed)	222-157-202-185-5	de–it
EUIPO – list of goods and services Spanish and English (Processed)	435-267-443-884-2	en–es
EUIPO – list of goods and services Spanish and French (Processed)	346-829-436-948-6	fr–es
EUIPO – list of goods and services French and English (Processed)	379-456-736-661-9	en–fr
EUIPO – list of goods and services Italian and English (Processed)	656-776-210-731-4	en–it
EUIPO – list of goods and services Italian and Spanish (Processed)	970-787-365-214-2	it–es
EUIPO – list of goods and services Italian and French (Processed)	885-745-599-446-6	it–fr
Bilingual resource with Bulgarian strategic documents in the field of innovations and digital growth (Bulgarian – English) (Processed)	224-075-564-720-7	en–bg
Parallel corpus from Estonian Cabinet of Ministers (Processed)	454-823-921-680-6	et–en
DA-EN Danish Ministry of Higher Education and Science 3 (Processed)	625-397-811-990-4	en–da
Translation memories from The Ministry of Foreign Affairs of Norway (Processed)	909-695-133-060-3	en–no
DA-EN Danish Ministry of Higher Education and Science 2 (Processed)	026-863-463-067-1	en–da
International Agreements (Processed)	810-722-062-476-6	lv–en
Parallel Corpus from the Web Site of the MFA of Latvia (Processed)	486-155-178-937-9	lv–en
Website of the President of the Republic of Lithuania (Processed)	967-335-099-703-2	en–lt
Bilingual documents Bulgarian-English in the field of transport (Processed)	539-070-524-117-8	en–bg
Parallel corpus from Bank of Estonia (Processed)	823-130-313-924-7	en–et
EJTN Handbook (Processed)	453-100-194-762-1	en–bg
Greek anti-corruption legislation and National Anti-Corruption Plan (Greek English) (Processed)	919-659-714-668-6	en–el
Translations of Lithuanian legislation from Seimas of the Republic of Lithuania (Processed)	691-158-541-313-8	en–lt
DA-EN Danish Ministry of Higher Education and Science (Processed)	222-781-852-505-9	en–da
Legal texts from Estonian Ministry of Justice (Processed)	556-164-094-707-8	en–et
Parallel corpus from Estonian Ministry of Foreign Affairs (Processed)	009-957-771-870-1	et–en
Corpus of State-related content from the Latvian Web (Processed)	636-211-843-827-4	lv–en
Romanian-English New Criminal Procedure Code (Processed)	085-350-774-090-4	en–ro
Bilingual resource with Bulgarian strategic documents in the field of telecommunications and broadband (Bulgarian–English) (Processed)	598-818-758-874-2	en–bg
DA-EN Danish Ministry of Higher Education and Science 4 (Processed)	560-401-490-272-1	en–da
Bilingual Bulgarian-English corpus from the National Revenue Agency (BG) (Processed)	039-753-902-920-1	en–bg
Central Statistical Office Dataset (Processed)	268-175-960-200-0	en–pl
PKN Orlen Dataset (Processed)	459-954-263-937-8	en–pl
Natolin European Centre Dataset (Processed)	238-889-529-582-8	en–pl
Polish Food Dataset (Processed)	136-690-396-444-3	en–pl
National Health Fund Dataset (Processed)	510-291-725-461-5	en–pl
Polish Food 4 & Food Policy Dataset (Processed)	801-617-774-005-7	en–pl
Polish Food Dataset 2 (Processed)	784-239-420-484-4	en–pl
Polish Food DataSet 3 (Processed)	871-723-818-068-0	en–pl
Polish Ministry of Foreign Affairs Regional Dataset (Processed)	957-995-509-395-3	en–pl
Polish Ministry of Foreign Affairs Historical Dataset (Processed)	615-782-338-621-1	en–pl

Polish Ministry of Foreign Affairs Youth 2011 Report (Processed)	910-694-644-926-4	en–pl
Public Procurement Dataset 2 (Processed)	865-835-648-658-1	en–pl
Civil Aviation Regulations (Processed)	792-786-685-848-5	en–pl
Public Procurement Dataset 1 (Processed)	141-723-057-887-8	en–pl
English-Slovak parallel corpus of texts from The Ministry of Culture of the Slovak Republic (Processed)	632-640-184-652-7	en–sk
English-Slovak parallel corpus of texts from The Ministry of Justice of the Slovak Republic (Processed)	734-218-150-302-0	en–sk
Secretariat-General parallel corpus SL-EN and EN-SL (part 1) (Processed)	271-870-307-699-4	sl–en
Secretariat-General parallel corpus SL-EN and EN-SL (part 2) (Processed)	963-471-195-725-8	sl–en
Romanian–English literature corpus (Processed)	050-476-818-226-7	en–ro
General Romanian-English bilingual corpus (Processed)	206-680-247-212-6	en–ro
Romanian–English news corpus (Processed)	100-905-126-706-7	en–ro
English-Norwegian parallel corpus from Forbruker Europa, 2017 release (Processed)	153-210-190-637-8	en–nb
Convention on the transfer of sentenced persons (English–Greek) (Processed)	114-726-489-848-3	en–el
BMVI Publications (Processed)	492-102-548-814-7	de–en
BMVI Website (Processed)	391-726-618-848-6	de–en
BMI Brochures and Website 2016 (Processed)	416-672-686-637-0	de–en
BMI Brochures 2011-2015 (Processed)	886-938-216-393-3	de–en
Luxembourg Museum Websites (de-en) (Processed)	280-308-415-749-7	de–en–fr
Parallel Global Voices (Greek–English) (Processed)	898-686-198-586-6	en–el
Parallel corpus (Greek–English) in the public administration domain (Processed)	530-246-044-238-5	en–el
Bilingual Croatian-English Parallel Corpus (Processed)	789-854-428-995-7	en–hr
Parallel corpus (Greek–English) in the law domain (Processed) (Part1)	923-676-825-761-4	en–el
Romanian Ombudsman archive (Processed)	422-693-047-625-3	en–ro
Macroeconomic Developments (Processed)	861-174-383-677-6	en–el
Methodological Reconciliation (Processed)	462-928-711-185-4	en–el
Expression of interest (Processed)	164-725-587-353-5	en–el
Memorandum for an ESM programme (Processed)	043-737-892-695-4	en–el
Parallel corpus (Bulgarian–English) in the public administration domain (Processed)	320-420-396-090-2	en–bg
Parallel corpus (Polish–English) from the website of the Polish Investment and Trade Agency (Processed)	898-973-186-773-1	en–pl
Parallel corpus from Social Insurance Agency – Försäkringskassan (Sweden) (Processed)	029-106-470-207-2	en–sv
English-Danish Parallel corpus from Tatoeba project (Processed)	893-698-207-679-6	en–da
Parallel corpus from Parliament of Estonia (Processed)	392-121-221-346-5	et–en
Corpus on Finance and Economics from Bank of Latvia (Processed)	389-271-130-137-6	lv–en
English-Finnish corpus from Finnish Information Bank (Processed)	894-719-306-863-7	en–fi
English-Estonian corpus from Finnish Information Bank (Processed)	492-203-674-156-9	et–en
English-Icelandic parallel corpus from Statistics Iceland (Processed)	968-796-585-795-9	en–is
Hallituskausi 2007-2011 – Finnish-English Translation Memory (Processed)	645-363-039-955-3	en–fi
Hallituskausi 2011-2015 – Finnish-English Translation Memory (Processed)	751-465-762-980-9	en–fi
English-Swedish corpus from Finnish Information Bank (Processed)	800-702-006-351-7	en–sv
The Gaois bilingual corpus of English-Irish legislation (Processed)	881-570-220-966-0	en–ga
The Coimisineir Teanga Bilingual Corpus of Reference Documents (Processed)	067-439-806-269-6	en–ga
Compendium The Social Insurance Institution (Processed)	736-818-016-630-9	en–pl
Bilingual hr-en parallel corpus from Croatian National Bank website (Processed)	248-991-649-363-5	en–hr
ENGLISH/POLISH PHRASE BOOK FOR ADMINISTRATIVE STAFF of LOCAL GOVERNMENT UNITS (Processed)	288-344-906-384-5	en–pl
Financial Stability Reports from the National Bank of Poland (2013-14) (Processed)	480-993-131-918-8	en–pl
Financial Stability Reports from the National Bank of Poland (2015-16) (Processed)	481-220-881-479-4	en–pl

The Coimisineir Teanga Bilingual Corpus of Reports and Press Releases (Processed)	440-182-838-797-0	en–ga
Parallel texts from the Swedish Competition Authority-Konkurrensverket (Processed)	450-184-406-181-3	en–sv
Maltese-English website parallel corpus (Processed)	693-091-524-649-2	en–mt
Malta Government Gazette (Processed)	032-192-530-108-8	en–mt
Laws of Malta (Processed)	833-404-387-881-6	en–mt
Polish Ministry of Foreign Affairs reports in EN and PL (Processed)	428-240-825-532-5	en–pl
Translation memory from Swedish National Audit Office (NAO) – Riksrevisionen (Processed)	709-518-556-855-4	en–sv
English-Swedish parallel corpus from texts of the Swedish Crime Victim Compensation and Support Authority (Brottsoffermyndigheten) web site (Processed)	520-443-837-112-0	en–sv
English-Croatian parallel corpus from texts of the Swedish Crime Victim Compensation and Support Authority (Brottsoffermyndigheten) web site (Processed)	659-571-021-533-3	en–hr
English-Swedish parallel corpus from the web site of the Swedish Migration Board – Migrationsverket (Processed)	689-938-422-892-7	en–sv
English-Swedish parallel texts from The Swedish Agency for Economic and Regional Growth – Tillväxtverket (Processed)	638-105-678-437-3	en–sv
Parallel Global Voices (English – Polish) (Processed)	993-121-813-887-5	en–pl
Employment in Poland 2009 report in EN-PL (Processed)	062-316-276-801-8	en–pl
Quarterly Reports of the Parliamentary Budget Office (Hellenic Parliament) (Processed)	497-530-909-088-2	en–el
Bilingual collection of reports of the Greek Public Power Corporation (Processed)	456-799-985-207-6	en–el
Portuguese-English bilingual corpus from Legislation concerning the Portuguese Parliament (Processed)	619-032-712-383-8	en–pt
Portuguese-English bilingual corpus from the Portuguese Constitution (Processed)	512-646-970-643-2	en–pt
Parallel corpus (en-pl) from the Export Promotion Portal of Poland (Processed)	545-583-786-881-8	en–pl
Spanish-English website parallel corpus (Processed)	664-503-904-200-9	en–es
Parallel corpus from the website of the Chancellery of the Prime Minister of Poland (Processed)	738-143-451-715-5	en–pl
Polish-English parallel corpus from the website of the Ministry of National Defence (Processed)	394-336-995-019-0	en–pl
Polish-English parallel corpus from the website of the Citizens Information Board (Processed)	247-240-766-663-4	en–pl
Polish-English parallel corpus from the website of the Ministry of Agriculture and Rural Development (Processed)	262-130-262-084-4	en–pl
Polish-English parallel corpus from the website of the Ministry of Development (Processed)	267-187-874-661-8	en–pl
Polish-English parallel corpus from the website of the Ministry of Justice (Processed)	273-889-173-953-2	en–pl
Polish-English parallel corpus from the website of the Ministry of Digitization (Processed)	637-353-316-483-2	en–pl
Polish-English parallel corpus from the website of the Ministry of Foreign Affairs (Processed)	445-973-164-231-9	en–pl
Polish-English parallel corpus from the website of the Ministry of Culture and National Heritage (Processed)	277-850-352-583-8	en–pl
Polish-English parallel corpus from the website of the Ministry of the Interior and Administration (Processed)	650-911-174-028-8	en–pl
Polish-English parallel corpus from the website of Public Employment Services in Poland (member of EURES network) (Processed)	693-193-256-174-1	en–pl
Polish-English parallel corpus from the website of the National Science Centre (Processed)	895-946-002-460-7	en–pl

Polish-English parallel corpus from the website of the ING Polish Art Foundation (Processed)	169-269-556-805-3	en–pl
Polish-English parallel corpus from the website of the National Security Bureau (Processed)	952-099-838-624-2	en–pl
Bilingual Bulgarian-English corpus from the 2018 Proposal for a National Climate Change Adaptation Strategy and Action Plan from the website of the Bulgarian Ministry of Environment and Water (Processed)	182-772-814-980-2	en–bg
Croatian-English corpus with statistical reports and studies from the Croatian Bureau of Statistics website (Processed)	378-424-378-060-6	en–hr
English-Estonian Parallel corpus compiled from translated annual reports from Estonian Academy of Sciences	841-714-744-394-2	et–en
Croatian-English corpus with studies on the challenges to the Croatian Accession to the European Union from the Croatian Institute of Public Finance website (Processed)	389-289-275-352-6	en–hr
Slovenian-English corpus with statistical reports from the Statistical Office of the Republic of Slovenia website (Processed)	169-569-336-630-0	sl–en
English-Swedish parallel corpus from Annual Reports of the Swedish Pension System (Processed)	747-260-595-720-9	en–sv
English-Swedish parallel corpus from the Annual Overview of Sweden's Official aid Agency SIDA Activities (Processed)	875-433-307-071-2	en–sv
Romanian-English corpus with studies, reports and statistical data in the field of culture from the National Institute for Cultural Research and Training website (Processed)	131-157-185-289-5	en–ro
Hellenic Ministry of Foreign Affairs Greek-English announcements corpus (Processed)	243-990-404-547-2	en–el
Greek-English parallel corpus from the website of the Prime Minister of the Hellenic Republic (Processed)	763-048-196-707-6	en–el
Bilingual hr-en parallel corpus from the Journal of the Croatian Association of Civil Engineers website (Processed)	732-156-538-451-4	en–hr
Polish-English parallel corpus from the website Business in Poland (Processed)	020-049-902-880-1	en–pl
Polish-English parallel corpus from the website Polish Aid (Processed)	441-363-587-087-7	en–pl
Polish-English parallel corpus from the website of the Polish Tourism Organisation (Processed)	042-465-256-319-2	en – pl
Polish-English parallel corpus from the website of the U.S. EMBASSY and CONSULATE IN POLAND (Processed)	770-000-350-631-0	en – pl
Polish-English parallel corpus from the website of the National Centre for Nuclear Research (Processed)	669-675-072-617-5	en – pl
Polish-English parallel corpus from the website of the Central Statistical Office (Processed)	715-697-723-261-3	en–pl
Polish-English parallel corpus from the website of the National Centre for Research and Development (Processed)	211-004-223-002-2	en–pl
Polish-English parallel corpus from the website of the Office of the Commissioner for Human Rights (Processed)	179-180-789-338-7	en–pl
Polish-English parallel corpus from the website of the Ministry of Regional Development (Processed)	912-804-204-923-3	en–pl
Polish-English parallel corpus from the website of the Institute of Mathematics of the Polish Academy of Sciences (Processed)	691-582-610-574-6	en–pl
Polish-English parallel corpus from the website of the Ministry of Digital Affairs (Processed)	600-324-789-103-0	en–pl
Polish-English parallel corpus from the website geoportal.gov.pl (Processed)	095-692-852-936-6	en–pl
Polish-English parallel corpus from the website of the Ministry of Science and Higher Education (Processed)	926-751-981-923-4	en–pl
Polish-English parallel corpus from the website Science in Poland (Processed)	962-404-942-173-7	en–pl
Polish-English parallel corpus from the website of the State Marine Accident Investigation Commission (Processed)	853-165-780-341-2	en–pl

Polish-English parallel corpus from the website of the National Audiovisual Institute (Processed)	382-527-820-478-0	en–pl
Polish-English parallel corpus from the website of the National Digital Archives (Processed)	536-218-040-486-1	en–pl
Croatian-English parallel corpus from the website of the Government Office for Cooperation with NGOs (Processed)	288-712-830-695-6	en–hr
Croatian-English parallel corpus from the website of the Embassy of Finland, Zagreb (Processed)	938-270-896-641-7	en–hr
Croatian-English parallel corpus from the website of the Ministry of Foreign and European Affairs, Republic of Croatia (Processed)	677-053-215-388-2	en–hr
Croatian-English parallel corpus from the website of the Croatian Journal of Fisheries (Processed)	403-261-656-321-0	en–hr
Croatian-English corpus with the Rural Development Programme for the Period 2014-2020 from the Croatian Rural Development Programme website (Processed)	994-259-799-079-0	en–hr
The Croatian-English corpus with the nature protection strategy of Croatia (Processed)	250-662-686-256-3	en–hr
Parallel Global Voices (Bulgarian – English) (Processed)	447-555-810-536-7	en–bg
Corpus of Icelandic texts from the Central Bank of Iceland (Processed)	420-670-865-427-1	is
Monolingual documents from the Government of Lithuania (Processed)	268-109-862-136-1	lt
Parallel texts from Swedish Labour market agency. Part 2 (Processed)	949-454-272-492-9	de–en–fi–es–pl–fr–sv–ro
Letter of rights for persons arrested on the basis of a European Arrest Warrant (Processed)	175-028-844-014-3	bg–it–de–en–lv–nl–fr–pl–el–ro
Parallel texts from Swedish Labour market agency (Processed)	496-669-153-886-4	de–en–fi–es–fr–sv–ro
Parallel texts from Swedish Social Security Authority (Processed)	002-471-002-734-6	it–de–en–fi–hr–es–pl–fr–sv–ro
Parallel texts from Swedish Work environment Authority (Processed)	448-438-055-941-1	bg–cs–it–de–en–fi–lv–hu–es–sv–et–pl–fr–lt–el–ro
Parallel texts from Swedish National Food Agency (Processed)	017-195-587-556-2	en–fi–es–pl–fr–sv
SIP Publications (Processed)	356-910-738-191-0	de–en–fr
Trilingual Documents related to International Judicial Cooperation in Civil Matters (Greek-English-French) (Processed)	954-287-236-137-4	en–fr–el
Letter of rights for persons arrested and or detained (Processed)	604-574-272-897-7	bg–en–lv–pl–fr–el–ro
Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment – United Nations (French-English-Greek) (Processed)	290-988-825-874-8	en–fr–el

Table 6: ELRC-SHARE – LRs provided to ELG

The summary of the 27 languages covered by the 187 LRs is shown in Table 7. Most of the LRs contain data in English, together with other languages.

1.	bg	13
2.	cs	3
3.	da	6
4.	de	17

5.	el	19
6.	en	179
7.	es	11
8.	et	9
9.	fi	8
10.	fr	17
11.	ga	3
12.	hr	16
13.	hu	1
14.	is	2
15.	it	8
16.	lt	5
17.	lv	8
18.	mt	3
19.	nb	1
20.	nl	1
21.	no	1
22.	pl	60
23.	pt	2
24.	ro	12
25.	sk	4
26.	sl	3
27.	sv	15

Table 7: ELRC-SHARE – languages covered by ingested datasets

Regarding the ELRA Catalogue, 22 LRs have been provided for these releases. Despite the fact that all the 103 pre-selected LRs were pre-converted to ELG metadata schema, only 22 could be handled at this stage due to the required sharing conditions. The specifics of this conversion will be detailed in Section 5, and, in particular, the licensing and sharing requirements in Section 6.2. Still, it should be mentioned that the use of ELRA LRs is defined by three parameters: “purpose of use”, “type of institution” and “membership”, which are currently covered by the ELG metadata schema but the imposed restrictions cannot be enforced in this release (e.g., there is no mechanism for identifying ELRA members). This is part of the future metadata extension plans (cf. Section 7.2). Furthermore, pricing and billing are functionalities to be defined and developed within the ELG platform, and which are required by many ELRA resources considering the three mentioned parameters.

The 22 LRs that have been imported into ELG can be seen in Table 8. These LRs are distributed under CC licenses and have been made available for download both from ELG and from the ELRA website. The latter approach was an intermediate measure while finalising the ELG imports for R1.

Resource name	**ISLRN**	**Languages**
CEPLEXicon	408-817-203-152-3	pt
MCL: Multifunctional Computational Lexicon of Contemporary Portuguese	489-956-642-755-8	pt
Translanguage English Database (TED) Transcripts database	502-719-830-448-5	en
Spanish Festival HTS models – male speech	017-935-913-932-1	es
Spanish Festival HTS models – female speech	653-517-560-115-8	es
Bilingual (Spanish-English) Speech synthesis HTS models	277-380-359-561-3	en–es
JV_TDM Corpus	371-240-320-910-4	fr

Speaking atlas of the regional languages of France	112-393-061-014-3	co–ca–rom–eu–nl–fr–br
GeFRePaC – German French Reciprocal Parallel Corpus	086-761-267-762-3	de–fr
PANACEA English-French and English-Greek parallel corpus acquired for Environment domain	870-946-931-293-7	en–fr–el
PANACEA English-French and English-Greek parallel corpus acquired for Labour Legislation domain	428-891-110-719-1	en–fr–el
PTPARL Corpus	294-303-577-819-2	pt
Nepali Monolingual written corpus	325-796-965-405-9	ne
English-Nepali Parallel Corpus	853-487-663-161-6	ne–en
Khresmoi manually annotated reference corpus	764-036-829-417-7	en
deL1L2IM corpus	339-799-085-669-8	de
2006 CoNLL Shared Task – Ten Languages	578-227-532-044-0	bg–de–pt–da–es–ja–sl–nl–tr–sv
NPChunks	412-883-442-173-8	pt
2007 CoNLL Shared Task – Basque, Catalan, Czech & Turkish	769-620-932-723-2	tr–cs–eu–ca
2007 CoNLL Shared Task – Greek, Hungarian & Italian	270-733-242-642-3	hu–it–el
Normalized Arabic Fragments for Inestimable Stemming (NAFIS)	305-450-745-774-1	ar
ECPC Corpus (European Comparable and Parallel Corpora of Parliamentary Speeches Archive) – set 1	036-939-425-010-1	en–es

Table 8: ELRA – LRs provided to ELG

The 23 languages represented in these 22 LRs are listed in Table 9.

1.	ar	1
2.	bg	1
3.	br	1
4.	ca	2
5.	co	1
6.	cs	1
7.	da	1
8.	de	3
9.	el	3
10.	en	7
11.	es	5
12.	eu	2
13.	fr	5
14.	hu	1
15.	it	1
16.	ja	1
17.	ne	2
18.	nl	2
19.	pt	5
20.	rom	1
21.	sl	1
22.	sv	1
23.	tr	2

Table 9: ELRA – languages covered by ingested datasets

The pre-selection of LRs for ingestion from the META-SHARE nodes has also been reduced due to licensing restrictions. The LRs that have been left aside for the time being comprise resources distributed under restrictive

META-SHARE proprietary licenses (such as MS-C-NoReD, MS-NC-NoReD and MS-Commons-BY-SA)[23] as well as other licenses that need to be negotiated with the data providers. The full list of imported LRs can be found in Table 10. The remaining resources to be negotiated will be part of the negotiation work to be done towards R2.

Only DFKI, ELDA and ILSP's META-SHARE nodes were considered for R1a and R1 due to licensing restrictions with the other nodes. The others are being approached as we write this deliverable, with the aim of establishing a collaboration for R2. A coordinated and scaled integration is planned, to optimise conversion and ingestion efforts by starting with work on the main managing nodes and following with all the others afterwards.

Resource Name	Language(s)
DeepBankDE	German
English ontology lexicon	English
eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing	English, German, Italian
Finance domain ontology	English
Finance English corpus	English
Finance English grammar	English
Finance English web corpus, automatically harvested	English
Greek Textual Entailment Corpus	Greek
Greek-Bulgarian Bul-TM parallel corpus	Bulgarian, Greek
ILSP PsychoLinguistic Resource	Greek
INTERA Corpus – the Bulgarian POS annotated part of the BG-EN pair	Bulgarian
INTERA Corpus – the Bulgarian structurally annotated part of the BG-EN pair	Bulgarian
INTERA Corpus – the Bulgarian-English part	Bulgarian, English
INTERA Corpus – the Bulgarian-English terms from the BG-EN pair	Bulgarian, English
INTERA Corpus – the English POS annotated part of the BG-EN pair	English
INTERA Corpus – the English POS annotated part of the EL-EN pair	English
INTERA Corpus – the English POS annotated part of the English-Slovene SVEZ ACQUIS Corpus	English
INTERA Corpus – the English POS annotated part of the SR-EN pair	English
INTERA Corpus – the English structurally annotated part of the BG-EN pair	English
INTERA Corpus – the English structurally annotated part of the EL-EN pair	English
INTERA Corpus – the English structurally annotated part of the EN-SL SVEZ ACQUIS corpus	English
INTERA Corpus – the English structurally annotated part of the SR-EN pair	English
INTERA Corpus – the English-Slovene terms from the EN-SL SVEZ ACQUIS Corpus	English, Slovenian
INTERA Corpus – the Greek POS annotated part of the EL-EN pair	Greek
INTERA Corpus – the Greek structurally annotated part of the EL-EN pair	Greek
INTERA Corpus – the Greek-English part	English, Greek
INTERA Corpus – the Greek-English terms from the EL-EN pair	English, Greek
INTERA Corpus – the Serbian POS annotated part of the SR-EN pair	Serbian
INTERA Corpus – the Serbian structurally annotated part of the SR-EN pair	Serbian
INTERA Corpus – the Serbian-English part	English, Serbian
INTERA corpus – the Serbian-English terms from the SR-EN pair	English, Serbian
INTERA Corpus – the Slovene structurally annotated part of the EN-SL SVEZ ACQUIS corpus	Slovenian
INTERA Corpus – the Slovene SVEZ ACQUIS POS annotated part of the EN-SL SVEZ ACQUIS Corpus	Slovenian
INTERA English-Slovene SVEZ ACQUIS Corpus	English, Slovenian
IT helpdesk Italian web corpus, manually harvested	Italian
IT helpdesk Spanish web corpus, automatically harvested	Spanish

[23] For more details on the META-SHARE licensing documentation http://www.meta-share.org/p/82/Legal-issues#licensing_scheme

IWSLT 2015 Human Post-Editing data	English, German, Vietnamese
IWSLT 2016 Human Post-Editing data	English, French, German
IWSLT 2017 Human Post-Editing data	Dutch, German, Italian, Romanian
KELLY word-list Greek	Greek
PANACEA Environment Corpus n-grams EL (Greek)	Greek
PANACEA Labour Legislation Corpus n-grams EL (Greek)	Greek
Parallel Global Voices	Albanian, Amharic, Arabic, Aymara, Bengali, Bulgarian, Burmese, Catalan, Chinese, Czech, Danish, Dutch, English, Esperanto, Filipino, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Khmer, Korean, Macedonian, Malagasy, Oriya, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Spanish, Swahili, Swedish, Turkish, Urdu
POETICON Multisensory and Multimedia Recordings of Everyday Interaction	English
SemEval-2016 ABSA Museum Reviews-French: Test Data-GOLD (Subtask 3)	French
SemEval-2016 ABSA Museum Reviews-French: Test Data-Phase A (Subtask 3)	French
SemEval-2016 ABSA Museum Reviews-French: Test Data-Phase B (Subtask 3)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-GOLD (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-Phase A (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-Phase B (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Train Data (Subtask 1)	French
TermCymru	English, Welsh
The TaraXÜ Corpus of Human-Annotated Machine Translations – 4rth evaluation round	Czech, English, French, German, Spanish
Tourism English grammar	English
Tourism Italian grammar	Italian
Travel domain ontology	English, German, Greek, Italian, Spanish
Travel English crowdsourced corpus	English
Travel English grammar	English
Travel English ontology lexicon	English
Travel English web corpus, automatically harvested	English
Travel English web corpus, manually harvested	English
Travel German ontology lexicon	German
Travel Greek grammar	Greek
Travel Greek web corpus, automatically harvested	Greek
WMT 2015 Human Evaluations	Not specified
WMT 2015 Translation Task Submissions	Czech, English, Finnish, French, German, Russian
WMT 2016 Human Evaluations	Not specified
WMT 2016 Translation Task Submissions	Czech, English, Finnish, German, Romanian, Russian, Turkish
WMT 2017 Human Evaluations	Not specified
WMT 2017 Translation Task Submissions	Chinese, Czech, English, Finnish, German, Latvian, Russian, Turkish
WMT18 Quality Estimation Task: Product Reviews	English, French

Table 10: META-SHARE – LRs provided to ELG

Table 11 provides an overview of the language coverage achieved with the META-SHARE LRs ingested.

1.	am	1
2.	ar	1
3.	ay	1
4.	bg	6
5.	bn	1
6.	ca	1
7.	cs	5
8.	cy	1
9.	da	1
10.	de	12
11.	el	19
12.	en	47
13.	eo	1
14.	es	4
15.	fa	1
16.	fi	3
17.	fil	1
18.	fr	12
19.	he	1
20.	hi	1
21.	hu	1
22.	id	1
23.	it	6
24.	ja	1
25.	km	1
26.	ko	1
27.	lv	1
28.	mg	1
29.	mk	1
30.	my	1
31.	nl	2
32.	or	1
33.	pl	1
34.	pt	1
35.	ro	3
36.	ru	4
37.	sl	4
38.	sq	1
39.	sr	5
40.	sv	1
41.	sw	1
42.	tr	3
43.	ur	1
44.	vi	1
45.	zh	3

Table 11: META-SHARE – languages covered by ingested datasets

Work towards R1 is concluded by end of April 2020. Given the results and needs derived from the R1a release, a strategic decision had been taken which involved focusing efforts on achieving a robust and stable version with R1, polishing and improving remaining technical and metadata issues from R1a before ingesting any further LRs.

With regards to the potential resources deriving from the ELG pilot projects, we are eagerly awaiting the evaluation results of the pilot project proposals in early summer 2020. As the first open call is still ongoing during the time of writing, the ingestion of such resources will be part of the work for R2 at the earliest. Once the call is closed and the submitted proposals evaluated, we will have a clearer picture of what to expect and what the pilot projects may be requiring. The type of resources to be treated (data resources, services, applications) will depend on the proposals we receive and their evaluation.

5 Language Resource Metadata Conversion

For R1, LRs from the following repositories were integrated into ELG: ELRA catalogue, META-SHARE repositories from DFKI, ELDA and ILSP nodes and the ELRC-SHARE repository.

To enable their insertion into ELG, the specificities of each repository, including ELG's, had to be taken into consideration and conversions in distinct steps were carried out, as presented below. The ELG metadata schema v1.0.2 is partly inspired from that of META-SHARE and has been adapted to ELG requirements. The ELG metadata model XSD has been made available on Gitlab[24] and it has been used as a reference in the metadata mappings. In order to optimise conversion, converters have been developed in a modular manner, which has allowed to reuse some components from one conversion to another. This is further explained below.

5.1 Conversion from ELRA Catalogue into ELG

To convert the Language Resources described in the ELRA Catalogue into the ELG metadata format, the steps described in the following subsections had to be taken care of.

5.1.1 Updating of ELRA Catalogue XML Schema Definition (XSD)

The ELRA Catalogue of Language Resources is a derived version of the META-SHARE structure which has been adapted to ELRA's specific distribution requirements. Before proceeding with the metadata conversion, ELDA had to provide consolidated XML files with respect to the META-SHARE metadata v3.1. To validate existing XML files from the ELRA Catalogue, we first made an analysis of discrepancies between the META-SHARE XSD and the current ELRA Catalogue XML files. From this analysis, a number of elements were detected that had to be corrected in the ELRA Catalogue XSD (see Annex A.B. for a list of reported errors). The updating of the ELRA Catalogue XSD was then done with the help of a Python script implemented at ELDA. The XML files could then be exported into META-SHARE 3.1 format.

5.1.2 Mapping between META-SHARE 3.1 and ELG-SHARE 1.0.2.

Once exported from the ELRA Catalogue into META-SHARE 3.1 format, a mapping had to be carried out between the ELRA XML files and the ELG metadata 1.0.2. This was implemented by ELDA. As for the previous step, where we had compared ELRA metadata with META-SHARE metadata, the mapping allowed us to adapt the

[24] https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema

validated ELRA Catalogue XML files (exported in META-SHARE 3.1 format) and make them compliant with ELG-SHARE. The list of elements that had to be adapted is given in Annex A.C.

The main issues that had to be addressed are summarised here with details outlined in Annex A.B. Examples of the elements that had to be reviewed and updated comprise the identifier (both format and content), some keywords that are attached to the resources have been added (using the ELRA values), some additional licenses have been added to include the ones used by ELRA, the "organization" element (part of the LRs catalogue) has been linked to the entity in the "Players/organization" database to be used by ELG, etc.

5.1.3 Conversion from META-SHARE 3.1 to ELG Metadata Model 1.0.2.

Once the mapping between the ELRA Catalogue and ELG was done, the conversion from META-SHARE 3.1 to the ELG model 1.0.2. could be implemented. This was done using XSLT.

5.2 Conversion from META-SHARE Repositories into ELG

META-SHARE's DKFI, ELDA and ILSP nodes are based on META-SHARE XSD 3.0. An XSLT was provided by ILSP to enable the conversion from META-SHARE XSD 3.0 to 3.1. In order to convert those META-SHARE nodes into ELG metadata 1.0.2, it was decided to start by converting the XML files from META-SHARE 3.0 into META-SHARE 3.1 with the already existing XSLT. Then, the XSLT produced in Section 6.1.3 was used to convert META-SHARE 3.1 XML files into ELG metadata 1.0.2 (as it had been done for the ELRA-SHARE conversion into ELG). This modular approach allowed to use META-SHARE v3.1 as pivot schema for conversion, reusing the implemented XSLTs for further conversions (such as ELRC-SHARE's in the following section).

5.3 Conversion from ELRC-SHARE Repository into ELG

The ELRC-SHARE repository is also a derived version of META-SHARE. To benefit from the existing ELRA to ELG meta-data converter, ELDA developed a Python script to first convert ELRC-SHARE resources in XML format into the ELRA Catalogue format, as an intermediary step. This ELRC-SHARE to ELRA script was built thanks to a mapping between a list of 187 XML files obtained from ELRC-SHARE and the ELRA Catalogue schema. To ensure consistency between the different metadata schemas, a number of changes and adaptations have been implemented, some are exemplified herein:

Changing from "underReview" value as "available" value in "distributionInfo" component.

- Redirecting "dataFormat" into "mimeType" in "textFormatInfo" component.
- Creating a default email address when these do not exist in the "communicationInfo" component (compulsory field in ELRC-SHARE).
- If field "validated" is present in the "validationInfo" component: insertion of "true" value.

As a complement to the metadata conversion, all ELRC-SHARE datasets associated to each metadata record were also added as compressed ZIP files in the ELRA Catalogue and their download access activated to ensure the usability of the converter from ELRA to ELG described in section 6.1.

Once the 187 ELRC-SHARE LRs had been converted into the ELRA format (based on META-SHARE), the same conversion methodology was applied as for the ELRA resources described in Section 5.1.

5.4 Insertion into ELG and Cleaning

The converted XML files from the different repositories were then added to the API developed for this purpose within ELG. However, due to some inconsistencies found while including them in the ELG API, some additional

modifications had to be done manually or semi-automatically in the metadata content. These were applied directly in the XML files in order to avoid further conversion developments (see all modifications reported in Annex A.D). For resources from the ELRA Catalogue, the corresponding modifications were also corrected in parallel in the ELRA Catalogue.

6 Identified Gaps

This section describes the gaps identified during the mapping and conversion work done for R1a and R1 and that need to be addressed within the project. These gaps are of a three-fold nature: technical, legal and financial/distributional.

6.1 Technical Issues

The ELG catalogue can be populated in the following ways:

- manual creation of metadata records using the metadata editor, or through the upload of an XML file,
- conversion of existing metadata records from their current schema and import into ELG.

In both cases, in order to be successfully imported and stored in ELG, a metadata record must comply with at least the **minimal version** of the schema. The definition of a minimal version is crucial for the population of the catalogue with a considerable number of useful resources. The full model is rich in information, capturing aspects of the whole lifecycle of a Language Resource/Technology. However, this makes the manual creation of metadata records a rather tedious and demanding task. In addition, the ingestion of LRTs from catalogues or repositories where they have been described with less informative models is hindered if adherence to the full model is a pre-requisite. Moreover, the quality and informativeness of the metadata records must be safeguarded. The minimal version sets a threshold of required information without which the metadata record becomes useless for ELG purposes (e.g., it could not be discovered by prospective users, or not properly used). Nevertheless, it should be pointed out that, while putting emphasis on fulfilling the minimal version requirements, we have not neglected other metadata elements for the conversion process. In fact, we have tried to convert as much metadata as possible from the original schemas into ELG, taking advantage of the **full version** of the model.

This is the main principle upon which the conversion process has been based. As described in previous sections, for this release, the process we followed involved:

- manual selection of resources from the three catalogues according to a set of pre-defined criteria,
- analysis of the source metadata schemas and mapping to the target one,
- implementation of conversion software,
- running of the converters on the selected resources,
- manual checks and import into a local database of the converted resources, identification of errors and mismatches,
- fixes of the identified gaps and errors through:
    - improvement of the conversion software and re-running of the software,
    - amendments of the metadata records on the source catalogues,
    - additional conversion scripts created and run on the converted resources,
    - and/or manual fixes of the converted resources.

Figure 4 provides a visual representation of the conversion and validation process in ELG. The whole process has been a time-consuming effort that took place in multiple rounds. Even though the model is implemented in the form of an XSD schema, which ensures that XSD validators can be (and have been) used to check that the converted metadata records adhere to the ELG schema, this has not been sufficient in identifying all issues.

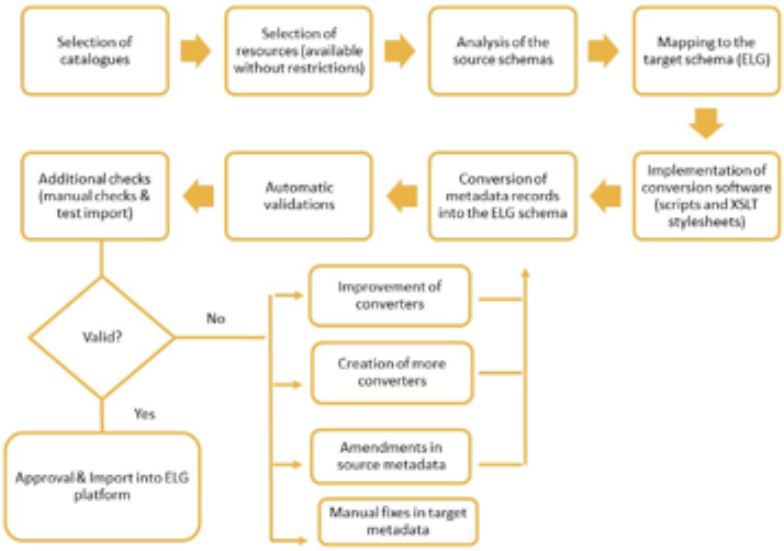

Figure 4: Conversion and validation process

The import into a local database has contributed to this, as the script used for uploading the XML files includes a set of rules that check for syntactic and partial semantic integrity. Such rules check for instance, the presence or value of a metadata element based on the value of another element, e.g., if the element “lingualityType” has a value “bilingual” or “multilingual”, the element “multilingualitytype” must also be filled in.

In addition, random sample manual checks have brought our attention to a further set of issues. For instance, the “language” element in ELG follows the BCP47 recommendation, which includes the ISO 639 for languages[25]. In the ELRA catalogue, for some resources, the “language” value was split into “language” and “languageVariety”. For instance, the ISO 639-1 code “es” corresponds to the value “Spanish; Castilian”, which in ELRA appeared as “language”: “Spanish” and “languageVariety”: “Castilian”. It is better that similar issues are reported to the providers and changed in the source catalogue before being converted. If for whatever reason this is not done on the provider’s catalogue, they have to be changed on the target ELG XML file.

At the final stage of the conversion process, we have also taken additional measures to overcome issues that result from the fact that not all functionalities are available for the current release of the platform. One such example relates to the deduplication of related entities, such as licenses, organisations, projects, etc. For these, the deduplication mechanism in release R1a relied on the internally assigned ELG identifier; thus, in order to avoid the proliferation of persons, organisations, etc. in the database, we have decided to first create metadata records for the related entities, and then add, with a script, their ELG identifier in the metadata records of the LRs and LTs where they are mentioned. For release 1, we have already implemented a mechanism that takes into account an element, such as email for persons, website for organisations and projects, etc., that can uniquely identify them; thus, this step is no longer required.

[25] Details about language coding: https://en.wikipedia.org/wiki/IETF_language_tag

As described in Section 3.3, the ELG metadata model is based on the META-SHARE model, on which the three catalogues that we have exploited for this population stage are also based but using different versions and different extensions. Because of this, the mappings have not always been straightforward. The main mismatches stem from the following.

- Elements that have become mandatory in ELG are not filled in the source metadata records, e.g., "format" for data resources. For this release, we had the benefit that the source catalogues are curated by consortium partners and, thus, could add the missing elements. For other catalogues, we will need to adopt a different strategy (see Section 6.2).
- Structural changes of the schema, i.e., elements that have been moved to another path. The most characteristic example for this is that of "format" and "size": In META-SHARE v3.0, these elements were attached to the resource; in ELG, they have been moved under "distribution", as they are properties that fit better the description of the physical form of a resource rather than the resource as a whole. Structural changes are implemented with the converters.
- Elements that took a "free text" value in the source schemas and are now filled in with a value from controlled vocabularies. These mappings and conversions have been made with some heuristics. Improvements for this type will be checked again in the following releases.

Finally, we should include here a technical issue that we identified and that is being addressed. This has to do with the upload of the data files by users. The upload mechanism has been implemented but due to technical reasons, it is not yet supported via a browser and, consequently, the catalogue User Interface. Therefore, currently, the uploading procedure is performed manually by system administrators, the data files are imported and stored on the storage system, and connected to the respective metadata records through the filename (common for the XML metadata file and the data file).

This first stage of conversion was highly instructive and very informative. The team will be using this valuable experience to speed up the processes of LR ingestion in the future releases. It is also important to mention that starting with some of the largest catalogues helped identify most of the issues that can be now anticipated. It is also clear that harvesting smaller repositories will be more straightforward.

6.2 Legal Issues

Managing legal issues in ELG is taking into account processes that may be different from one provider to another. In particular, a provider may choose to distribute language resources either through implicit or explicit licenses, through specific and controlled conditions of use as well considering a certain user status such as profiles or membership status. Such issues imply the setting up of a legal expert team.

6.2.1 Implicit versus Explicit Licenses

One main distinction that has to be taken into account is the management of implied (or implicit) versus express (or explicit) licenses. The definitions according to Wikipedia for the two license types are as follows:

Implied license[26]: An implied license is an unwritten license which permits a party (the licensee) to do something that would normally require the express permission of another party (the licensor). Implied licenses may arise by operation of law from actions by the licensor which lead the licensee to believe that it has the necessary permission.

[26] https://en.wikipedia.org/wiki/Implied_license

Express license[27]: The opposite of an implied license is an express license, which, for some forms of intellectual property, must be in writing.

For implied licenses, it has now become a commonly and widely used practice to grant users access when they click on the license terms acceptance button indicated on the repository pages.

6.2.2 Conditions of Use

The conditions of use[28] of a LR or LT are another factor that should be taken into account and which may require further discussion and interaction between the provider and the user.

Among the various elements to consider in licensing data or tools, we can, at this phase of the project (R1), simply mention that we will review the purpose of use (which could be for commercial purpose, for research, etc.), as well as the profile of the licensee (this is the type of institution, some resources may have be restricted to particular types of institutions, e.g., "academic" or "commercial"). Other elements will be instantiated in due time and according to the requirements and expectations of the ELG community.

6.3 Financial and Distribution Issues

Not only legal issues may condition the delivery of resources to a user but also the financial and distribution policies of the provider. Such policies involve a dedicated team, with expertise in technical, legal and financial domains. Some LRs may be available for free, whereas others will be available for a fee. Such information shall be clearly displayed to users. Some resources are available under different pricing terms:

- **Legal profile of the licensee**: prices may be different depending on the profile of the users and the license they apply for, e.g., a not-for-profit organisation can also opt for a commercial licence.
- **Purpose of use**: prices may depend on different scenarios of use, as detailed in section 6.2.2.
- **Pricing policy**: depending on the distribution policy, some discounts may be offered, either on a continuous or on a limited period of time. The pricing schema will be covering all possible scenarios that the ELG community may express (e.g., resource downloads versus temporary use within the platform).

7 Next Steps

7.1 Next Steps for the Ingestion into ELG

A number of points have already been raised along this deliverable as next steps in the ongoing work to enrich the ELG platform. These are planned for R2 and listed below.

[27] https://en.wikipedia.org/wiki/Implied_license#Express_license
[28] We use here "Conditions of use" for the LRs and LTs and can also read "Terms of use" (though these are more used in web sites).

Provision of LRs

- Remaining negotiations for META-SHARE licenses are envisaged to secure the addition of LRs that could not be ingested for R1a and R1 due to their licensing conditions.
- Further nodes from the META-SHARE network will be ingested into ELG following agreements with their respective node managers.
- Work on LINDAT's analysis and harvesting will be finalised for R2.
- Analysis of OLAC will be concluded to identify important repositories from all registered repositories.
- Analysis of the repository list identified by the ELG consortium (provided in Annex A.A) will be concluded with further prioritisation of repositories for R2 and R3.

Metadata

- ELRA resources restrictions (concerning ELRA membership) will be reviewed for R2 so as to include all available LRs from the ELRA Catalogue.
- For further points to be addressed relating to metadata from other repositories see Section 7.2.

Policies with an impact on metadata

- Billing and payment will be addressed for LRs/LTs, analysing their impact on the ELG metadata schema. This will be relevant once commercial products are made available through the ELG platform.

Technical issues

- Issues detected during the conversion of LRs and population of ELG R1 will be tackled (see Section 7.2).

Pilot projects

- The strategy of ingestion of LRs from ELG pilot projects will be designed, considering what could be their outcomes and accounting for their requirements and needs (in collaboration with the partners from CUNI, who are responsible for the Open Calls).

7.2 Plans for the Import of Metadata from Other Sources

The population of the ELG platform from catalogues that use a schema similar to ELG has given us the chance to identify issues (Section 6) and prepare for upcoming ELG releases. The steps to be followed, regarding the technical aspect of the conversion from other sources, fall into two groups.

Collection and improvement of current tools and contents: The process we have followed for the conversion and correction of the metadata records involved the implementation of various tools (scripts, XSLT stylesheets, etc.), with the META-SHARE v3.1 schema as the middle layer. The first steps to be taken, to the extent possible, will focus on collecting and merging the existing tools, consolidating, and documenting the issues they resolve, and enriching them with solutions to the already identified gaps.

Establishment of a formal procedure for mass population of the ELG platform from other sources and support with documentation and tools, including the following.

- Automatic and periodical metadata harvesting: In contrast to the current manual selection, conversion and upload of resources, this will ensure that we are always up-to-date with the contents of the source

catalogue/repository. For this step, we are already investigating the OAI-PMH protocol[29] for harvesting, but also other solutions, such as ResourceSync and DOIP, that will allow us also to import data together with the metadata (in accordance with the agreements with their right holders). It should be noted, though, that for providers that do not want to implement harvesting mechanisms, the current solution (manual harvesting and/or case-by-case ingestion) will remain possible.

- Guidelines for repositories as prospective providers: Irrespective of the solution to be adopted for the automatic harvesting, we will also need a clearly stated specification of the type of resources to be harvested (i.e., similar to the pre-selection procedure of this release) and instructions on the metadata to be exposed. Concerning the metadata schema, besides the full version that is already documented, the ELG guidelines (see D2.4) include a detailed description of a minimal version with explanations, instructions, and examples. These will be enriched with additional instructions for repositories, both for the completion of the metadata and its exposing.

The discussed issues with respect to the metadata schema may need to be addressed, including:

- Missing mandatory metadata elements: For those elements that do not create issues, a value such as "Not available" or "Undefined" could be devised. Where possible, alternatives for semi-automatically deriving values from other metadata elements may be discussed and put in place in collaboration with the providers. However, the absence of a value for elements deemed crucial for the completeness of the metadata record (e.g., media type) or the safe use of the resource (e.g., license) may lead to the rejection of a metadata record. Remedies will be discussed with the providers on a case-by-case basis.
- Mappings of free text values to controlled vocabularies or between different controlled vocabularies: Generic tools that can be used for such conversions (e.g., existing converters between the different ISO codes of languages, tools implementing similarity algorithms that can be used for matching free text values to values from controlled vocabularies, etc.) will be collected and shared. New tools that will be implemented for this task will also be shared as a source of inspiration for similar issues.
- Checking and curation of converted metadata records: To the extent possible, (semi-)automatic checks looking into the quality of the converted metadata record will be exploited to identify problems and report back to the providers. Depending on the issue, the appropriate measures will be implemented together with the providers.

[29] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), https://www.openarchives.org/pmh/

A. Annexes

A.A. Identified Inventories and Repositories

This table provides an overview of the work in progress regarding the identification of inventories and repositories as language resource sources. Together with the Name and URL of the source, a short preliminary description of their content is provided. The complete analysis of this list is foreseen for R2.

Name	URL	Short preliminary description
OPUS – the open parallel corpus	http://opus.nlpl.eu/	Mentioned in the document (multilingual LRs for MTs), many pairs
Tilde MODEL Corpus – Multilingual Open Data for European Language	https://tilde-model.s3-eu-west-1.amazonaws.com/Tilde_MODEL_Corpus.html	Mostly EU languages (often from all to Fr, En, Lv)
WMT resources	http://www.statmt.org/wmt19/translation-task.html	Mentioned in the document (WMT selects multiple language for each year), e.g., 2018 English-Chinese, English-Czech English-Finnish and Finnish-English English-German and German-English English-Gujarati and Gujarati-English English-Kazakh and Kazakh-English English-Lithuanian and Lithuanian-English English-Russian and Russian-English French-German and German-French And this year two low-resource language pairs (English to/from Kazakh and Gujarati)
CELR resources	https://keeleressursid.ee/en/resources	This is the centre for Estonian Languages, multiple resources related to Estonian (also the CLARIN local centre)
Corpora collected by University of Tarty	https://www.cl.ut.ee/korpused/	Focus on Estonian textual corpora (mostly in the CLARIN Estonian Centre)
Open Data portal of Republic of Latvia	https://data.gov.lv/eng/	The Latvia's Open Data portal (more general repository, may not be on a high priority for ELG)
Latvian National Terminology Portal	https://termini.gov.lv/	The national terminology portal of Latvia (Terminologies, mostly monolingual)
eTranslation TermBank	https://www.eurotermbank.com/	The European consolidated Terminology bank (almost 15M terms and 42 Languages), produced by multiple EU funded projects)
European Union Terminology	https://iate.europa.eu/download-iate	The official EU terminology, with almost 1M terms and 26 languages.
Språkbanken Text	https://spraakbanken.gu.se/eng/resources	Nationella språkbanken (the National Language Bank) of Sweden. Many resources related to Swedish with all modalities of interest to ELG.
ELRA Universal Catalogue	http://universal.elra.info/	An ELRA catalogue that stores resources not available for distribution yet, over a hundred of languages and various modalities (mostly resources identified during conferences)
LRE Map	http://lremap.elra.info/	Resources listed by the authors of papers submitted to LREC (over 5000 not-unique resources associated with publications)

META-SHARE Node: CNR — National Research Council of Italy	http://META-SHARE.ilc.cnr.it/	META-SHARE node
META-SHARE Node: FBK — Fondazione Bruno Kessler	http://META-SHARE.fbk.eu/	META-SHARE Node
META-SHARE Node: IPIPAN — Institute of Computer Science, Polish Academy of Sciences	http://META-SHARE.nlp.ipi-pan.waw.pl/META-SHARE/	META-SHARE Node
META-SHARE Node: Tilde	http://META-SHARE.tilde.com/	META-SHARE Node
META-SHARE Node: Budapesti Műszaki és Gazdaságtudományi Egyetem	http://META-SHARE.tmit.bme.hu/	META-SHARE Node
META-SHARE Node: Filozofski fakultet Sveučilišta u Zagrebu	http://meta-share.ffzg.hr/	META-SHARE Node
META-SHARE Node: Institute for Bulgarian Language, Bulgarian Academy of Sciences	http://META-SHARE.ibl.bas.bg/	META-SHARE Node
META-SHARE Node: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences	http://META-SHARE.korpus.sk/	META-SHARE Node
META-SHARE Node: Lithuanian Language Institute	http://meta-share.lki.lt/	META-SHARE Node
META-SHARE Node: National Library of Norway	http://META-SHARE.nb.no/	META-SHARE Node
META-SHARE Node: Research Institute for Linguistics, Hungarian Academy of Sciences	http://META-SHARE.nytud.hu/	META-SHARE Node
META-SHARE Node: Romanian Academy Center for Artificial Intelligence	http://ws.racai.ro:9191/	META-SHARE Node
META-SHARE Node: University of Belgrade	http://meta-net.matf.bg.ac.rs:8080/META-SHARE/	META-SHARE Node
META-SHARE Node: University of Copenhagen	http://META-SHARE.cst.dk/	META-SHARE Node
META-SHARE Node: University of Gothenburg	http://spraak-banken.gu.se/META-SHARE/	META-SHARE Node
META-SHARE Node: University of Helsinki	http://META-SHARE.csc.fi/	META-SHARE Node
META-SHARE Node: University of Lisbon	http://META-SHARE.meta-net4u.eu/	META-SHARE Node
META-SHARE Node: University of Lodz	http://META-SHARE.ia.uni.lodz.pl/	META-SHARE Node
META-SHARE Node: University of Tartu	http://META-SHARE.ut.ee/	META-SHARE Node

META-SHARE Node: University of the Basque Country	http://aholab.ehu.es/META-SHARE/	META-SHARE Node
META-SHARE Node: University of Vigo	http://META-SHARE.gts.uvigo.es/	META-SHARE Node
META-SHARE Node: Universitat Politècnica de Catalunya	http://META-SHARE.talp.cat/	META-SHARE Node
META-SHARE Node: Universitat Pompeu Fabra	http://META-SHARE.upf.edu/	META-SHARE Node
META-SHARE Node: Synthema	http://META-SHARE.synthema.it:8000/	META-SHARE Node
META-SHARE Node: University of Malta	http://META-SHARE.research.um.edu.mt/	META-SHARE Node
CLARIN Virtual Language Observatory	https://vlo.clarin.eu	The CLARIN catalogue, a large set of resources provided by the CLARIN Network (mostly for EU Languages, a particular focus on Social sciences and humanities)
Linguistic Data Consortium Catalogue	https://catalog.ldc.upenn.edu/	The LDC catalogue comprises over 500 resources, many funded by the US DARPA and covering over 100 languages, some of the major languages (English, Arabic, Chinese) but also minority ones from Africa and Asia.
SADiLaR Language Resource Repository	https://repo.sadilar.org/	The new repository for the south African languages (initial focus on the 12 official languages, both speech and texts, over 370 resources)
The Speech Ark	http://www.thespeechark.com/products.html	A very few speech resources (4 or 5) developed by The Speech Ark, a spin-out company of the University of Birmingham (British English)
Eurac Research CLARIN Centre	https://clarin.eurac.edu/	A CLARIN Centre
CLARIN-ERIC	https://www.clarin.eu/	The CLARIN Managing and coordinating centre, see VLO
AIDA Data Hub	https://datasets.aida.medtech4health.se/	Mostly Image resources (radiology area)
UCI Machine Learning Repository	http://archive.ics.uci.edu/ml/index.php	A Data Centre with very few textual corpora (about 40 but a lot of mathematical data etc.). Focus on multilingualism.
LINDAT/CLARIN repository	https://lindat.mff.cuni.cz/repository/xmlui/	A CLARIN Centre
ILC-CNR for CLARIN-IT repository	https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/	CLARIN and META-SHARE Centre in Italy
CLARIN-PL	https://clarin-pl.eu/dspace/	A CLARIN Centre
HunCLARIN	http://clarin.hu/content/hun-clarin-tagjai	A CLARIN Centre
CLAPOP	http://portal.clarin.nl/	A CLARIN Centre
CITK	https://toolkit.cit-ec.uni-bielefeld.de/	A few resources listed at the Bielefeld University (a couple of IR QA datasets)
SinMin	https://osf.io/a5quv/	The site contains texts of different genres and styles of the modern and old Sinhala language. Mostly textual corpora.

CLARIN-UK	https://www.clarin.ac.uk/	A CLARIN Centre
Repository CLARIN-D Centre Leipzig	http://clarin.informatik.uni-leipzig.de/repo/	A CLARIN Centre
CLARIN repository at the University of Tübingen	https://uni-tuebingen.de/en/134314	A CLARIN Centre
Aboriginal Studies Electronic Data Archive (ASEDA)	http://www1.aiatsis.gov.au/ASEDA/ASEDAsr.xml	Information about Aboriginal and Torres Strait Islander languages which has been assembled from a number of referenced sources. May be useful for TTS and Social sciences.
Archive of the Indigenous Languages of Latin America (AILLA)	http://www.ailla.utexas.org/	Mostly resources for linguistics and social sciences (small resources from Indigenous languages)
C'ek'aedi Hwnax Ahtna Regional Linguistic and Ethnographic Archive	https://scholarspace.manoa.hawaii.edu/handle/10125/4538	Hawaii university repository comprises a few resources (corpora) that could be repackaged for LTs.
Central Institute of Indian Languages: Publications	http://www.ciil.org/Main/Publications/publication.asp	A Central Institute of Indian Languages, a government entity with a few glossaries and dictionaries
COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)	http://cocoon.huma-num.fr/exist/crdo/	Part of the Huma-num, a very large archiving centre for resources produced in France both for technology and Humanities. A few hundreds of resources (mostly French)
The Crúbadán Project	http://crubadan.org/	A project to collect resources, nothing available as LRs
ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics “A. Zampolli”, National Research Council, in Pisa	http://www.clarin-it.it/	CLARIN Centre
Kaipuleohone	http://scholarspace.manoa.hawaii.edu/handle/10125/4250/	The Hawaii university repository of local languages comprises Audio Data and Transcriptions
Language Commons Language Corpora	http://www.archive.org/details/LanguageCommons	International consortium that is creating a large collection of written and spoken language material, made available under open licenses.
Language resources at the Text Laboratory	http://www.hf.uio.no/tekstlab/	A few resources at the university of Oslo
The LINGUIST List Language Resources	http://linguistlist.org/olac/	Not a repository per se but it disseminates announcements about resources and should be kept in our list to check regularly.
Magoria Books’ Carib and Romani Archive	http://archive.magoriabooks.com/	Some resources for Cariban and Romani languages (textual data, lexica, etc.)
Oxford Text Archive	http://ota.oucs.ox.ac.uk/	The Oxford Text Archive is one of the first repositories of full-text literary and linguistic resources. More than 25 languages are mentioned.
POLLEX-Online	http://pollex.org.nz	A very specialised repository with a large-scale comparative dictionary of Polynesian languages.
SAILS Online	http://sails.clld.org	Very few resources related to the South American Indigenous Languages

Slovenian language resource repository CLARIN.SI	http://www.clarin.si/	CLARIN Centre
Speech and Language Data Repository (SLDR/ORTO-LANG)	http://sldr.org	A French centre for Spoken resources (mostly French but also associated with over 20 languages from Europe, China, Africa)
TALKBANK Data repository	http://talkbank.org/	TalkBank is the result of the project organized by Brian MacWhinney at Carnegie Mellon University, Resources related to CHILDES Corpus (Clarin)
Webonary Sites	https://www.webonary.org	Part of SIL (coordinating OLAC and Ethnolog), this is a repository of Dictionaries and Grammars of the World
TROLLing	https://dataverse.no/dataverse/trolling	A CLARIN Center (DataverseNO) , worth mentioning here because of the focus on Arctic languages (DataverseNO is a national research data repository in Norway)
The Rosetta Project	http://rosettaproject.org/	Famous Rosetta Project with an ambitious project of the construction of a universal corpus of human language by collecting parallel text and audio in the world's 300 most widely-spoken languages
Comparalex	http://comparalex.org/index.php?page=about	A Canadian national institute with a database of language word list data with audio samples (Canadian and other associated languages from, e.g., French countries but not very LT oriented)
LAUDATIO	http://www.laudatio-repository.org/repository/	A cooperation project between INRIA and Humboldt-Universität zu Berlin (about 20 textual Corpora, German and some few other languages (English))
SAVEE	http://kahlan.eps.surrey.ac.uk/savee/	This is one database (but very interesting), The Surrey audio-visual Expressed Emotion (SAVEE) DB
Recola Database	https://diuf.unifr.ch/main/diva/recola/download.html	This is one important database: RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions
IEMOCAP DATABASE	https://sail.usc.edu/iemocap/	Another emotion database (The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multi-speaker database, recently collected at SAIL lab at USC.), US English.
datasets-CMU_Wilderness	https://github.com/festvox/datasets-CMU_Wilderness	CMU Wilderness Multilingual Speech Dataset: over 700 different languages providing audio, aligned text and word pronunciations of the Bible (http://www.bible.is/)
Spanish CLARIN K-Centre	http://www.clarin-es-lab.org/index-es.html	CLARIN Centre
Buckeye Speech Corpus	http://buckeyecorpus.osu.edu/	This is one resource, the Buckeye Corpus of conversational speech (high-quality recordings from 40 US speakers conversing freely with an interviewer). The speech has been orthographically transcribed and phonetically labelled.
Data and Service Center for the Humanities	http://data.dasch.swiss	The Data and Service Center for humanities (DaSCH), member of the Swiss Academy of Humanities and Social Sciences hosts a number of resources (old books /OCR data)
IMS Universität Stuttgart Repository	http://clarin04.ims.uni-stuttgart.de/repo/	CLARIN Centre

Kielipankki	https://www.kielipankki.fi/language-bank/	The Language Bank of Finland, with over 200 resources (mostly Nordic languages inc. Sami, almost half downloadable.
Pacific and Regional Archive for Digital Sources in Endangered Cultures	http://www.paradisec.org.au/	The focus is on endangered languages and PARADISEC mentions 500 collections representing over 1,200 languages (a collection could be images or audio recordings).
ARCHE	https://arche.acdh.oeaw.ac.at/browser/	A Resource Centre for Humanities Related Research in Austria with a dozen resources useful for LT, multiple languages (including English, Arabic)
Arquivo.pt – the Portuguese web-archive	http://www.arquivo.pt	Not a repository but a web archive that may be useful for future crawling.
Michigan Corpus of Academic Spoken English	http://quod.lib.umich.edu/m/micase/	The Michigan Corpus of Academic Spoken English
TextGrid Repository	https://textgridrep.org	The TextGrid Repository (TextGridRep) is a digital preservation archive for human sciences research data providing a variety of data for teaching and research purposes.
Repository CLARIN-D Centre CEDIFOR	https://www.cedifor.de/repository-clarin-d-centre-cedifor-2/	CLARIN Centre
Australian National Corpus	https://www.ausnc.org.au/	A small repository of Australian English (about twelve resources of Australian English text, transcriptions, audio and audio-visual materials).
CLARINO Bergen Center repository	https://repo.clarino.uib.no/xmlui/	CLARIN Centre
Deutsches Textarchiv	http://www.deutschestextarchiv.de/	CLARIN (associated) Centre
Huygens ING	https://www.huygens.knaw.nl/?lang=en	Repository of medieval digitalized resources (Dutch)
CLARIN-DK-UCPH Repository	https://repository.clarin.dk/repository/xmlui/	CLARIN Centre
English Lexicon Project	http://elexicon.wustl.edu/	Access to a large set of lexical characteristics, along with behavioural data from visual lexical decision and naming studies.
Tekstlaboratoriet	http://www.hf.uio.no/iln/english/about/organization/text-laboratory/	C Centre in the European CLARIN infrastructure.
Reading Experience Database	http://www.open.ac.uk/Arts/reading/	Lists of books no more copyrighted but could be crawled for corpus production
Speech and Language Data Repository	http://www.sldr.org	The French centre for speech and text deposit (SLDR/ORTOLANG, mostly French but also with multiple languages, e.g., Hindi, Marathi , Arabic, in addition to EU languages.
Language Archive Cologne	https://lac.uni-koeln.de	CLARIN Centre
CLARIN service center of the Zentrum Sprache at the BBAW	https://clarin.bbaw.de	CLARIN Centre
CLARIN INT Portal	https://portal.clarin.inl.nl/	CLARIN Centre
IDS Repository	http://repos.ids-mannheim.de/	CLARIN Centre
GAMS	http://gams.uni-graz.at/context:gams	Medieval resources (texts and OCR, multiple languages)

Datenbank Gesprochenes Deutsch	https://dgd.ids-mannheim.de/	Database for Spoken German (Datenbank für Gesprochenes Deutsch), a small number (Germans living outside Germany)
CLARIN-LT	http://clarin-lt.lt/?lang=en	CLARIN Centre
Centre National de Ressources Textuelles et Lexicales	http://www.cnrtl.fr/	Few French textual corpora, high quality annotations
Romani Morpho-Syntax Database	http://romani.humanities.manchester.ac.uk/rms/	A few resources of Romani
clarin:el inventory of language resources and services	https://inventory.clarin.gr/	CLARIN Centre
Bibliothèques Virtuelles Humanistes	http://www.bvh.univ-tours.fr/	Medieval OCRed resources (Textual and Old French)
CLARIN Centre Vienna	https://clarin.oeaw.ac.at/ccv/	CLARIN Centre
Informatics Research Data Repository	https://www.nii.ac.jp/dsc/idr/en/index.html	Part of the Japanese infrastructure for Data archiving and distribution, a few hundreds of resources (mostly Speech and Text in Japanese but also a couple of English and South-Asia resources).
Hamburger Zentrum für Sprachkorpora Korpus Repositorium	https://corpora.uni-hamburg.de/hzsk/en/repository-search	Mostly Spoken data for German (40) but also a few for Spanish, English, Old High German, or Polish.
Bavarian Archive for Speech Signals	https://clarin.phonetik.uni-muenchen.de/BASRepository/	The BAS repository contains corpora of spoken language archived in the Bavarian Archive for Speech Signals (BAS) at the university of Munich. Hundreds of resources (many available also via ELRA), mostly German and some English.
The Language Archive	https://tla.mpi.nl/	This is the Max Plank Institute (see if the resources mostly field-linguistics are still available)
Phonogrammarchiv	https://www.oeaw.ac.at/phonogrammarchiv/	The Austrian Academy of Sciences (also CLARIN Centre)
Meertens Instituut Collecties	http://www.meertens.knaw.nl/cms/en/	The Meertens Institute hosts a few resources of Dutch (lexical, named entities, etc.)
UdS Fedora Commons Repository	http://fedora.clarin-d.uni-saarland.de/index.en.html	CLARIN Centre

Table 12: Overview of currently examined inventories and repostories

A.B. Validation of ELRA Catalogue XML Files

As indicated in Section 5.1.2, the mapping between the ELRA XML files (exported from ELRA-SHARE to META-SHARE 3.1) and ELG-META-DATA 1.0.2) helped identify a number of metadata elements that had to be aligned. It was easy to convert and/or update the ELRA elements to ensure consistency with the ELG extensive metadata elements that capitalized on all recent experiences of the consortium (ELRA, META-SHARE, ELRC-SHARE, etc.). Some of these issues are described here and a detailed table is given at the end of this annex.

How to deal with some required information

- /MetadataRecordIdentifier: if the LR already has an LR identifier (e.g., ISLRN, handle PID), we use the element “LRIdentifier” with the relevant scheme on the attribute.

- /LanguageResource/keyword: at least one keyword is needed. For ELRA Catalogue resourceType value was chosen.
- /LexicalConceptualResourceTextPart/metalanguage: Currently optional and to be discussed for the next release R2. Currently, for ELRA Catalogue “und” (undetermined) value was chosen.
- /LanguageResource/LRIdentifier: This is about transforming ELRA IDs under /resourceInfo/identificationInfo/identifier into /LanguageResource/LRIdentifier with LRIdentifierScheme value: 'http://w3id.org/meta-share/meta-share/other'.

Information not having a direct mapping in ELG

- /LRIdentifier: /META-SHAREId (NOT_DEFINED_FOR_V2) not converted into /LRIdentifier
- /LicenceTerms/licenceTermsURL; each licence must have a unique name, url and id. We are using the SPDX list of (free and open source software/documentation) licenses for standard licenses[30]. However, for ELRA licenses that include different types of conditions, we had to create as many instances as possible conditions, i.e., 12 different “licenceTerms” (Table 13).

License name	restrictionsOfUse	userNature	Member
ELRA_END_USER	NONCOMMERCIALUSE	ACADEMIC	MEMBER
ELRA_END_USER	NONCOMMERCIALUSE	ACADEMIC	NOMEMBER
ELRA_END_USER	NONCOMMERCIALUSE	COMMERCIAL	MEMBER
ELRA_END_USER	NONCOMMERCIALUSE	COMMERCIAL	NOMEMBER
ELRA_VAR	COMMERCIALUSE	ACADEMIC	MEMBER
ELRA_VAR	COMMERCIALUSE	ACADEMIC	NOMEMBER
ELRA_VAR	COMMERCIALUSE	COMMERCIAL	MEMBER
ELRA_VAR	COMMERCIALUSE	COMMERCIAL	NOMEMBER
ELRA_EVALUATION	EVALUATIONUSE	ACADEMIC	MEMBER
ELRA_EVALUATION	EVALUATIONUSE	ACADEMIC	NOMEMBER
ELRA_EVALUATION	EVALUATIONUSE	COMMERCIAL	MEMBER
ELRA_EVALUATION	EVALUATIONUSE	COMMERCIAL	NOMEMBER

Table 13: ELRA license conditions

- /distributionInfo/availability: not mapped.
- /distributionInfo/licensor (ELG has a definition of licensor, but this is not mentioned on /LexicalConceptualResource/DatasetDistribution): not mapped.
- /distributionInfo/licenceInfo/restrictionsOfUse: see above decision for licenses.
- /distributionInfo/distributionRightsHolder/organizationInfo/communicationInfo/: organisations are a separate entity with their own metadata records. “templates” for organisations were added and shared with all partners.
- /distributionInfo/userNature/{academic|commercial}: see above decision for licenses.
- /distributionInfo/membershipInfo/member/{true|false}: see above decision for licenses.

[30] https://spdx.org/licenses/

Elements for which there exists a mapping, but in different components:

- /distributionInfo/distributionRightsHolder transferred into: /LanguageResource/LRSubclass/LexicalConceptualResource/DatasetDistribution/distributionRightsHolder. Extremely unlikely, but if needed, if there are two different distributionRightsHolders, e.g., for a downloadable form vs. the form of a lexicon accessed via an interface, we could add them in the two different distributions.
- /distributionInfo/iprHolder transferred into /LanguageResource/iprHolder.

The following summary shows the list of errors resulting from the comparison between the ELRA Catalogue XML files and the META-SHARE XSD.

attributionText- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}attributionText': This element is not expected. Expected is ({http://www.meta-share.org/META-SHARE_XMLSchema}licenceInfo).
audioFormatInfo- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}audioFormatInfo': This element is not expected. Expected is one of ({http://www.meta-share.org/META-SHARE_XMLSchema}languageInfo, {http://www.meta-share.org/META-SHARE_XMLSchema}modalityInfo, {http://www.meta-share.org/META-SHARE_XMLSchema}audioSizeInfo).
audioSizeInfo- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}audioSizeInfo': This element is not expected. Expected is ({http://www.meta-share.org/META-SHARE_XMLSchema}languageInfo).
corpusAudioInfo- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}corpusAudioInfo': Missing child element(s). Expected is one of ({http://www.meta-share.org/META-SHARE_XMLSchema}languageInfo, {http://www.meta-share.org/META-SHARE_XMLSchema}modalityInfo, {http://www.meta-share.org/META-SHARE_XMLSchema}audioSizeInfo).
distributionInfo- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}distributionInfo': Missing child element(s). Expected is ({http://www.meta-share.org/META-SHARE_XMLSchema}licenceInfo).
fundingType- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}fundingType': [facet 'enumeration'] The value 'serviceContract' is not an element of the set {'other', 'ownFunds', 'nationalFunds', 'euFunds'}.
fundingType- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}fundingType': 'serviceContract' is not a valid value of the local atomic type.
licence- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}licence': [facet 'enumeration'] The value 'CC-BY-4.0' is not an element of the set {'CC-BY', 'CC-BY-NC', 'CC-BY-NC-ND', 'CC-BY-NC-SA', 'CC-BY-ND', 'CC-BY-SA', 'CC-ZERO', 'PDDL', 'ODC-BY', 'ODbL', 'MS-NoReD', 'MS-NoReD-FF', 'MS-NoReD-ND', 'MS-NoReD-ND-FF', 'MS-NC-NoReD', 'MS-NC-NoReD-FF', 'MS-NC-NoReD-ND', 'MS-NC-NoReD-ND-FF', 'MSCommons-BY', 'MSCommons-BY-NC', 'MSCommons-BY-NC-ND', 'MSCommons-BY-NC-SA', 'MSCommons-BY-ND', 'MSCommons-BY-SA', 'ELRA_END_USER', 'ELRA_EVALUATION', 'ELRA_VAR', 'CLARIN_PUB', 'CLARIN_ACA', 'CLARIN_ACA-NC', 'CLARIN_RES', 'AGPL', 'ApacheLicence_2.0', 'BSD_4-clause', 'BSD_3-clause', 'FreeBSD', 'GFDL', 'GPL', 'LGPL', 'Princeton_Wordnet', 'proprietary', 'underNegotiation', 'nonStandardLicenceTerms'}.
licence- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}licence': 'CC-BY-4.0' is not a valid value of the local atomic type.
licence- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}licence': 'publicDomain' is not a valid value of the local atomic type.
metadataInfo- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}metadataInfo': This element is not expected. Expected is one of ({http://www.meta-share.org/META-SHARE_XMLSchema}distributionInfo, {http://www.meta-share.org/META-SHARE_XMLSchema}contactPerson).
sizeUnit- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}sizeUnit': [facet 'enumeration'] The value 'translationUnits' is not an element of the set {'terms', 'entries', 'turns', 'utterances', 'articles', 'files', 'items', 'seconds', 'elements', 'units', 'minutes', 'hours', 'texts', 'sentences', 'bytes', 'tokens', 'words', 'keywords', 'idiomaticExpressions', 'neologisms', 'multiWordUnits', 'expressions', 'synsets', 'classes', 'concepts', 'lexicalTypes', 'phoneticUnits', 'syntacticUnits', 'semanticUnits', 'predicates', 'phonemes', 'diphones', 'T-HPairs', 'syllables', 'frames', 'images', 'kb', 'mb', 'gb', 'rb', 'shots', 'unigrams', 'bigrams', 'trigrams', '4-grams', '5-grams', 'rules', 'questions', 'other'}.
sizeUnit- Schemas validity error : Element '{http://www.meta-share.org/META-SHARE_XMLSchema}sizeUnit': 'translationUnits' is not a valid value of the local atomic type.

A.C. List of Elements Mapped to ELG Metadata Schema 1.0.2

actualUseDetails
annotatedElements
annotationMode
annotationModeDetails
annotationType
audioGenre
availabilityStartDate
byteOrder
creationEndDate
creationMode
description
distributionAccessMedium
domain
downloadLocation
durationUnit
email
encodingLevel
extratextualInformation
fee
givenName
identifier
iprHolder
ISLRN
languageId
languageScript
languageVarietyName
languageVarietyType
lastDateUpdated
lexicalConceptualResourceType
licence
lingualityType
linguisticInformation
mediaType
membershipInstitution
metadataCreationDate
metadataLastDateUpdated
mimeType
multilingualityType
multilingualityTypeDetails
nonStandardLicenceTermsURL
organizationName
projectName
quantization
recordingDeviceTypeDetails
recordingEnvironment
region
relationType
resourceName
resourceShortName
resourceType
revision
samplingRate
scenarioType
segmentationLevel
signalEncoding
signConvention
size
sizeUnit
sourceChannel
sourceChannelDetails
speechItems
surname
targetResourceNameURI
timeCoverage
title
url
useNLPSpecific
validated

A.D. Modifications in Final Converted XML Files

Manual or semi-automatic modifications were made for the following fields.

For all resources

- addition of prefix “ms”
- addition of languageId for metalanguage, language
- change of xml:lang=“en-us” into “en”; check and make appropriate changes for xml:lang=“und” (the rule asks for at least an English value when the element is multilingual)
- removal of <annotation> element (made optional in the XSD for raw corpora)
- addition of website for ELRA
- only one datasetDistribution was kept when the licencesTerms was one of the CC licences (redundant as all other elements are exactly the same)
- downloadLocation replaced with accessLocation (according to DCAT: downloadLocation is used only for direct download locations, i.e., where there’s no extra click needed)

Changes were also made specifically for this release

- surname-name were reversed as they were occasionally wrong
- addition of ELG IDs for related entities (persons, organizations, licenses) (temporary measure)

1. **For resources from ELRA Catalogue**
   - correction of lingualityType
   - correction of mutilingualityTypeDetails
   - addition of multilingualityType when not present (mandatory when lingualityType=multilingual)
   - correction of language
   - correction of languageVariety
   - addition and correction of size
   - correction of dataFormat

2. **For resources from ELRC-SHARE**
   - change annotationType “domainSpecificAnnotation” to “alignment” (for ELRC, this was done automatically as we already knew the type of processing performed on these resources)
   - removal of second licenceTermsName for specific licences (e.g., openUnderPSI): multilingual elements are multiple in the XSD but only one value per language is allowed when imported
   - removal of double organizationName [rule: multilingual elements are multiple in the XSD but only one value per language is allowed when imported]
   - removed double projectName: multilingual elements are multiple in the XSD but only one value per language is allowed when imported
   - ELG identifiers were ignored (they are automatically assigned when imported in the database)
   - Creation of a HTML page for publicDomain and openUnderPSI, addition as value for licenceTermsURL
   - removal of languageVariety when not needed, e.g., the ISO code for “es” stands for “Spanish; Castilian”; languageVariety for Castilian is invalid in this case
   - emails were separated with commas to follow validation pattern for email
   - change resourceName of relatedLR to the resourceName without the “(processed)” mention: this was done automatically and may include errors compared to the original ELRC metadata record, but it was preferred over the resource id number