Chapter 8
Datasets, Corpora and other Language Resources

Victoria Arranz, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Penny Labropoulou, Miltos Deligiannis, Leon Voukoutis, and Stelios Piperidis

Abstract This chapter provides an overview of what is available in ELG in terms of datasets, corpora and other language resources (LRs) and how this has been achieved. We look at the procedures and steps that have been followed to complete the full resource ingestion cycle, which goes from repository and LR identification to metadata description and ingestion. We explain the approaches, priorities and methodology. The chapter also outlines the repositories that have been integrated into ELG, discussing the different procedures followed (metadata conversion, extraction, and completion, as well as harvesting) and the reasons behind these choices. Furthermore, the ELG catalogue content is described, with details on key elements and features as well as accomplishments. The last two sections are devoted to the crucial legal issues behind such a complex platform and its data management plan, respectively.

1 Introduction

As introduced in Part I, one of the ELG platform's primary functions is enabling sharing, distribution and deployment of Language Resources and Technologies (LRT). ELG provides access to thousands of datasets, by far the largest collection of relevant datasets for the European Language Technology community. Users can search for, download as well as provide different types of resources. As can be seen further down, ELG has identified, filtered, described and centralised a vast amount of datasets and other resources from different inventories and repositories, providing an easy to use point of search for the LT community. Its aim is to become the "yellow pages" and the primary platform for the European Language Technology community (see Chapter 9). Our work in terms of curating and further enriching ELG is ongoing, with new ingestions and collaborations at the time of writing.

Victoria Arranz · Khalid Choukri · Valérie Mapelli · Mickaël Rigault
ELDA, France, arranz@elda.org, choukri@elda.org, mapelli@elda.org, mickael@elda.org

Penny Labropoulou · Miltos Deligiannis · Leon Voukoutis · Stelios Piperidis
Institute for Language and Speech Processing, R. C. "Athena", Greece,
penny@athenarc.gr, mdel@athenarc.gr, leon.voukoutis@athenarc.gr, spip@athenarc.gr

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_8

This chapter describes the work carried out so far as well as currently ongoing efforts towards the population of the ELG catalogue with Language Resources (datasets and language models). This work has consisted in 1. the identification of sources (inventories and repositories), language resources and models, 2. their analysis, 3. the selection of elements to be ingested, as well as 4. the conversion or harvesting of their metadata descriptions and 5. the ingestion of these descriptions, and actual LRs, if relevant. All these steps are complex and intertwined tasks that are operationalised in a collaborative manner.

As a core element of ELG, the term "Language Resource" (LR, LRs) is used for resources composed of linguistic material used in the development, 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. In related initiatives and the literature, the term is often used with a broader meaning, encompassing also tools and services used for the processing and management of datasets, and standards, guidelines and similar documents that support the research, development and evaluation of LTs. In the ELG metadata model (see Labropoulou et al. 2020, and also Chapter 2), we use the term as first defined for the META-SHARE metadata model (Gavrilidou et al. 2012), i. e., including both data resources and LT tools/services. The alternative term Language Resource/Technology (LRT) is also used in the context of ELG (Rehm et al. 2021). However, in this chapter we use LR as referring to datasets and language models only; tools and services in ELG are discussed in Chapter 7.

2 Identification of Language Resources and Repositories

ELG aims to become the primary marketplace for the European LT community. The organisations making use of it range from commercial to non-commercial, including 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 helping them make their tools, services and data available and also establishing the means for them to find and have access to those they may require in their work.

To cover all relevant existing language resource repositories, ELG defined an identification and collection methodology. First, the ELG project consortium members performed a round of identification and analysis contributing their own resources. Second, we reached out to the ELG National Competence Centres (NCCs, see Chapter 11) to gather more input and pointers to additional existing repositories and resource inventories. This identification task has been run in parallel with a priority definition task, which has been adjusted regularly according to achievements and to the community's needs and demands.

2.1 Identification by the Consortium

ELG examined the available inventories and repositories of all potential LT/LR providers and users. The initial results have been completed with further collaborative input from the NCCs (see Section 2.2) and ELG's sister project European Language Equality (ELE, see Section 2.3.2). With regard to the typology of LRs searched for, all types and modalities deemed useful for some sort of LT application were considered. These comprise corpora, lexicons, terminologies, and derived resources (such as language models for ASR or TMX models for MT), and also focus on media such as speech/audio, text, video/audio-visual, images, OCR and sign language datasets (images, videos). The identification strategy was adjusted following initial findings. For example, users' needs guided us to take into account high-priority dataset types such as language models, and has led us to look into repositories which contain and even focus on such types of resources (see Section 4.2).

2.2 Identification by the National Competence Centres

In addition to the work described above (Section 2.1), a survey was carried out to gather more input from the NCCs and from other collaborators, often related to their local and regional repositories (Rehm and Marheinecke 2019). This way we have been able to identify new repositories and, moreover, we were also provided with extensive documentation by the NCCs (content, contacts, etc.). The collaboration with the NCCs has been valuable. We plan to continue the joint work to maximise ELG's coverage.

2.3 Collaboratively Filling the Gaps

With its (at the time of writing) 8,873 dataset descriptions and following the ingestion of several repositories, ELG is at a compelling stage for taking the next steps in its dataset provision strategy. It must be stressed that our collaboration with other initiatives has also had an impact on these numbers. Bearing that in mind, the population of ELG now follows the analysis and identification of gaps from several perspectives:

1. The ELG consortium members' analysis of contributions and ingestion statistics in the platform.
2. The analysis of gaps carried out under a joint strategy, such as the ELE project and the ELG pilot projects (see Part IV), which have contributed datasets and also shared their own needs with regard to ELG, thus supporting ELG on its LRT collection venture from the point of view of the provider and the user.

3. The analysis of feedback received from technology developers and data users who shared their needs with us.

2.3.1 Contributions from the ELG Pilot Projects

The ELG pilot projects were intended to demonstrate the usefulness of ELG by contributing datasets or services to the platform or by making use of existing datasets or services for the development of innovative LT applications. These contributions provided by the pilot projects benefit both the community that will have access to the assets provided as well as the pilot projects themselves that will gain visibility with their work and by displaying it in ELG. These projects are an excellent proof of concept for the ELG platform and those pilot projects that provide datasets often target – and fill – specific gaps. At the time of writing, the already concluded pilot projects have finished their work, which has resulted in a set of 52 datasets available through ELG. The pilot projects are described in detail in Part IV of this book.

2.3.2 Contributions from the European Language Equality Project

ELG collaborates with the European Language Equality (ELE) project[1] to promote digital language equality in Europe. In 2021, ELE organised an online survey addressed primarily to the more than 30 language experts of the consortium to collect information on language resources and technologies available for the languages[2] under investigation (see Chapter 6 for more details). Through a web form, the ELE consortium partners responsible for one or more of the languages addressed by the project were able to record and report new language resources and also new resource repositories. This additional and collaborative collection procedure resulted in approx. 6,300 records (Arranz et al. 2022), which have already been cleaned up, normalised and curated and finally ingested into ELG (4,127 metadata records for data resources and 2,215 metadata records for tools). Just like ELG organisation pages, metadata records can be claimed by the resource creators or other rightful owners (see Chapter 9, Section 3.3, p. 179) and enriched with further information. This is why all contact persons included in these metadata records have been notified of their publication in ELG; we encouraged them to claim their resources and enrich the descriptions. Complete metadata descriptions are an important aspect of ensuring findability and future reuse of the resources (see Chapter 2, Section 7).

[1] https://european-language-equality.eu

[2] https://european-language-equality.eu/languages/

2.3.3 Platform Users

Finally, users of the ELG platform can also provide feedback about their interaction with ELG or about unmet expectations with regard to the availability of datasets or LT services. With regard to the latter, if users raise a certain need for specific datasets in relation to specific technologies, the ELG team can investigate whether relevant datasets or resources exist.

3 Integrating Repositories into ELG

The individual ELG releases follow an evolutionary strategy with regard to the population of the catalogue. This strategy has evolved as procedures have been put in place and new priorities and needs identified. ELG Release 1 (R1) followed a rather pragmatic approach, exploring procedures while targeting large repositories under the management of ELG consortium members. This allowed us to set up procedures, locate flaws and address problems (e. g., pending legal issues). ELG Release 2 (R2) launched an ambitious acquisition of very large catalogues which were not compliant with ELG's structure and metadata schema. This was the case, for instance, for Quantum Stat and Zenodo (see Section 4 and Arranz et al. 2021). Repositories like Zenodo are extremely large digital libraries in which many different research artefacts are published, which is why it requires a certain amount of effort to find and extract artefacts that are relevant for ELG. Despite these challenges, the overall result is rewarding as it provides access to many LT-related datasets, which have not been directly discoverable so far and which are now made available to the community through ELG as a one-stop-shop. The LR provision strategy for ELG Release 3 (R3) has built on top of the processes firmly established in R2. It continued and finished up the integration of the already initiated repositories, it set up harvesting procedures for as many ingested repositories as possible and added further repositories.

3.1 Priorities in the Ingestion Work

The list of identified repositories comprised different types of portals, such as those storing data from evaluation campaigns or shared tasks (e. g., WMT resources, Yeganova et al. 2021), large catalogues of language resources (e. g., ELRA, Mapelli et al. 2022), networks of LR repositories (e. g., various META-SHARE nodes, Piperidis et al. 2014), databanks, initiatives supporting the collection of language data, etc. This initial list was prioritised by taking into account the following dimensions of the different repositories:

- Relevance of their content for ELG, its services and users.
- Access information (conditions of use, prioritising open licensing schemes).
- Languages covered (covering multiple different languages, filling detected gaps).

- LR typology (covering different modalities, filling detected gaps).
- Number of resources (prioritising repositories with larger numbers of resources).
- Metadata schema (prioritising schemas that allow automated conversions).

Following this prioritisation strategy, three repositories – all of which are run by members of the ELG project consortium – were initially selected for ingestion in ELG Release 1: ELRA[3], ELRC-SHARE[4] and the three META-SHARE nodes managed by DFKI[5], ELDA[6] and ILSP[7]. This choice was strategic, as a proof of concept for resource availability and metadata conversion, given that the involved partners were familiar with the content and metadata schemas of these repositories. All the datasets selected for metadata ingestion were filtered down for legal compliance to ensure that licensing or distribution conditions that could not be addressed by ELG at this early stage could be taken care of for a later release. ELG Release 2 continued with additional repositories under the management of ELG project consortium partners (ELRA-SHARE-LRs 2014, 2016, 2018 and 2020[8], and LINDAT/CLARIAH-CZ[9]) but also by extending its work on the META-SHARE network and looking into very large digital inventories such as Quantum Stat and Zenodo. The reasons behind these choices combined strategy and diversity, which were also the goal with repositories such as Hugging Face for ELG Release 3 (see Section 4.2.4).

3.2 Contributing Language Resources

Interested institutions or individuals can make datasets available for download, i. e., hosting datasets in the ELG platform, or they can simply point ELG users to external download locations. In both cases, a description of the resource in the form of a metadata record is needed that can be discovered through the ELG catalogue. Such metadata descriptions can be manually created in ELG using the corresponding editor, they can be prepared as an XML file, which is then uploaded and imported into ELG, or they can be automatically converted from existing metadata records that use a different schema and imported into ELG afterwards. The flexibility behind these different options to populate the ELG catalogue makes contributions very easy, they can be done according to the provider's needs and preferences.

ELRC-SHARE follows the metadata-only option; this repository is financed by the European Commission under the ELRC initiative (Lösch et al. 2021), datasets will be available through ELRC-SHARE for at least the duration of the ELRC contracts. For that reason, the master copies of the LRs provided to ELG remain within

[3] http://catalogue.elra.info

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

[5] http://metashare.dfki.de

[6] http://metashare.elda.org

[7] http://metashare.ilsp.gr:8080

[8] LRs contributed by LREC participants, see http://www.elra.info/en/lrec/shared-lrs/.

[9] LINDAT is the CLARIN Centre for Language Research Infrastructure in the Czech Republic.

ELRC-SHARE but corresponding metadata records are available through ELG, enabling their discovery through ELG and their download via a redirect to the corresponding ELRC-SHARE page. In addition to contractual reasons, some repositories prefer to host their LRTs themselves, such as the ELRA catalogue, which distributes its LRs under a typology of licences that cannot be fully covered or recreated by the ELG metadata schema for the time being. Repositories like Zenodo or Quantum Stat mostly provide links to the locations of their datasets, very often these are links to Github or Gitlab pages. Again, only metadata records with the links to the dataset locations have been ingested into ELG. Likewise, harvested repositories only export metadata records (e. g., different CLARIN nodes or Hugging Face).

4 Procedures to Ingest Language Resources

Different repositories need to be approached differently with the goal of extracting metadata records and ingesting them into ELG. This relates to a number of dimensions that have allowed us to categorise repositories and, thus, to set up procedures to process them. These relate to the *conversion*, *extraction and completion* as well as *harvesting* of LR metadata, further described in Sections 4.1, 4.2 and 4.3 below.

4.1 Metadata Conversion

We converted (through mapping) the metadata records of several repositories so that we could import them into the ELG catalogue, which follows the ELG metadata schema (Labropoulou et al. 2020). This was the case for the ELRA catalogue, the META-SHARE nodes and the initial ingestion of the ELRC-SHARE repository (managed through harvesting now, see below). This conversion work is complex, but it has paved the way for improvements and updates on both sides of the conversion line, on both the source and target metadata elements and descriptions.

4.1.1 From ELRA Catalogue to ELG

The conversion of the LR metadata entries in the ELRA catalogue into the ELG metadata format followed several steps:

- *Updating the ELRA catalogue XML Schema Definition (XSD):* The ELRA catalogue is based upon the META-SHARE structure, it has been adapted to ELRA's specific distribution requirements. Before proceeding with the metadata conversion, an analysis of discrepancies between the META-SHARE XSD and the ELRA catalogue XML files was performed. This allowed us to update the ELRA catalogue XSD and to export the XML files in META-SHARE 3.1 format.

- *Mapping between META-SHARE 3.1 and ELG-SHARE 1.0.2:* Once exported, the ELRA XML files were mapped to the ELG metadata schema 1.0.2. This mapping allowed us to adapt the validated ELRA XML files (in META-SHARE 3.1 format) and to make them compliant with the ELG-SHARE model. Several elements had to be adapted for that purpose.
- *Conversion from META-SHARE 3.1 to ELG Metadata Model 1.0.2:* Once the mapping between the ELRA catalogue and ELG was completed, we implemented an XSLT stylesheet to transform the META-SHARE 3.1 format to the ELG metadata model.

While the implementation of this first tool required quite a bit of effort, the experience gained was valuable for the subsequent implementation of other converters.

4.1.2 From META-SHARE to ELG

META-SHARE's DKFI, ELDA and ILSP nodes are based on META-SHARE XSD 3.0. An already existing XSLT stylesheet was used to convert from META-SHARE XSD 3.0 to 3.1. We implemented a second XSLT stylesheet to convert META-SHARE 3.1 XML files into ELG metadata 1.0.2 (as for the ELRA-SHARE conversion into ELG). This modular approach allowed us to use META-SHARE v3.1 as pivot schema, reusing the implemented XSLTs stylesheets for further conversions (such as ELRC-SHARE's below).

4.1.3 From ELRC-SHARE to ELG

ELRC-SHARE is also based on META-SHARE. The initial ingestion was carried out through conversion, a harvesting protocol was put in place later (see Section 4.3 and Chapter 6 in Part I). To benefit from the ELRA to ELG metadata converter, a subset of ELRC-SHARE LRs was converted first into the ELRA and then into the ELG format.

4.1.4 Import into ELG

The XML files converted from the metadata of the different repositories were then imported into ELG using the API developed for this purpose. Some inconsistencies remained that led to corrections both in the XML files and the ELRA catalogue.

4.2 Metadata Extraction and Completion

Now we look into those repositories that did not allow for a straightforward conversion or for which building converters was not a feasible option.

4.2.1 Zenodo

Zenodo[10] is a digital library launched in May 2013 within the OpenAire[11] project, to enable the compilation of research artefacts, such as publications, images, datasets, software, etc. A good number of those artefacts consists of LRs that may be of interest to the LT community. However, the extremely high number of artefacts in Zenodo together with the incompatibility of the Zenodo and ELG metadata schemes made the identification of relevant LRs a big challenge. We opted for a semi-automatic approach to collect what ELG considers as LRs, using a combination of Python and directly querying the Zenodo database, among others.[12] However, the compilation of metadata information still required manual intervention to ingest our selection of actual LRs as well as to add the minimal set of metadata elements which are mandatory for ELG and which do not exist in the Zenodo records. This semi-automated process required a lot of manual effort. We currently work on an automated harvesting-oriented approach (see Section 4.3 and Chapter 6 in Part I).

4.2.2 ELRA-SHARE-LRs

The ELRA-SHARE-LRs are provided by participants attending the Language Resources and Evaluation Conference (LREC). Participants can share the LRs they present at the conference either by uploading them in a special LREC repository or by linking them to their original download location using an online form. We selected a subset of these LRs by checking the compliance of licences with the ones accepted in ELG. Licences that are too vague were left aside (e. g., “Open Source”, “Creative Commons” without further specification). Given that the original metadata was available as a spreadsheet, the sheet and conversion tool produced to gather Zenodo metadata (see above) was adapted. As the ELRA-SHARE-LRs metadata contained only a minimal set of information, missing but required information was added manually into the spreadsheet to comply with the mandatory ELG metadata (e. g., type of LR, linguality, annotation, data format, licence, etc.). Finally, the spreadsheet was converted into XML and ingested into ELG.

4.2.3 Quantum Stat

Quantum Stat enables LR producers to register datasets in the “Big Bad NLP Database”.[13] The procedure for identifying, describing and ingesting datasets into ELG is as follows: first, an initial table with 481 datasets was exported and analysed for relevance to ELG by checking licensing information (whether licences are well

[10] https://zenodo.org

[11] https://www.openaire.eu

[12] https://developers.zenodo.org/#records

[13] https://datasets.quantumstat.com

identified), dataset type, and whether the resource can be downloaded. The datasets not complying to the LR description requirements were discarded and only compliant metadata information was kept. Then, as for ELRA-SHARE-LRs and Zenodo, the minimal set of metadata information was compiled, while also adding missing information before the actual conversion into XML and ingestion into ELG.

4.2.4 Hugging Face

Often described as a "model zoo", the Hugging Face[14] repository includes a large collection of machine learning models and datasets that can be used for training new models, with a focus on the Transformers architecture (Wolf et al. 2020). ELG collaborates with Hugging Face regarding the import of Hugging Face metadata records into ELG. One challenge relates to the fact that the description of resources in Hugging Face does not follow a specific methodology. To begin with, adding descriptions to resources is encouraged but not mandatory. Furthermore, the suggested metadata elements do not follow a standard schema. The manual work needed to process the filtered entries was considerable in order to enrich the information available. A conversion process was applied based on mapping the elements (see Chapter 6 for more details).

4.3 Metadata Harvesting

We implemented metadata harvesting solutions for ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN-PL and CLARIN-SI as well as Zenodo, as described below.

4.3.1 ELRC-SHARE

Three groups of datasets were originally selected from the three prioritised repositories to be converted and ingested into ELG Release 1 (see Section 4.1). Of these, only ELRC-SHARE allowed for the import of the whole list given that its resources met the following conditions: their licensing conditions allowed it (all data were shared under CC-BY licences, they were open under the directive on the re-use of public sector information, or they belong to the public domain), and their metadata elements were compatible and fully covered by the ELG metadata schema. We have implemented an OAI-PMH[15] client that harvests metadata records compliant with the ELG metadata schema, and we use this for regular harvesting from ELRC-SHARE.

[14] https://huggingface.co

[15] Open Archives Initiative Protocol for Metadata Harvesting (2015).

4.3.2 LINDAT/CLARIAH-CZ

The LINDAT/CLARIAH-CZ repository makes its metadata available for harvesting through its OAI-PMH end-point.[16] Means for ingesting metadata complying to the META-SHARE schema[17] were already in place in ELG and the repository did provide a mapping from its internal metadata storage to META-SHARE. An attempt was made at reusing this conversion, but the result was deemed unacceptable as not all of the available metadata was mapped. After a few iterations we arrived at a mapping between concepts that are important and required in the ELG schema and the metadata stored in LINDAT/CLARIAH-CZ. LINDAT updated the metadata for several of its resources following the feedback received from ELG. Also, based on the feedback from LINDAT/CLARIAH-CZ, some changes were applied to the ELG schema. The implementation of this mapping represents around 1,200 changed lines of code, including some tooling to reflect some of the metadata issues discovered.[18]

4.3.3 CLARIN-PL and CLARIN-SI

The LINDAT/CLARIAH-CZ repository makes available an OAI-PMH endpoint which exposes ELG-compatible metadata records. The repository software developed by the LINDAT/CLARIAH-CZ team, based on DSpace, is also used by several other CLARIN centres for their repositories, i. e., their metadata records are ready to be imported into ELG using the same harvesting procedure. For ELG Release 3, this collaboration has resulted in the regular harvesting of the CLARIN centres in Slovenia (CLARIN-SI) and Poland (CLARIN-PL).[19]

4.3.4 Zenodo

As described in Chapter 6 (Part I), Zenodo is a particularly interesting catalogue for ELG purposes. Zenodo exposes its metadata records through a REST API[20] as JSON data and through an OAI-PMH API[21] in a set of standard metadata formats, i. e., DC[22], DataCite[23], MARC21[24] and DCAT[25]. Work is currently ongoing to replace the semi-manual import of Zenodo metadata records that started for ELG Release

[16] http://lindat.mff.cuni.cz/repository/oai/request?verb=Identify

[17] http://www.meta-share.org/p/93/Documentation

[18] https://github.com/ufal/clarin-dspace/pull/930

[19] http://www.clarin.si and https://clarin-pl.eu

[20] https://developers.zenodo.org/#rest-api

[21] https://developers.zenodo.org/#oai-pmh

[22] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[23] https://schema.datacite.org/meta/kernel-4.4/

[24] https://www.loc.gov/marc/bibliographic/

[25] https://www.w3.org/TR/vocab-dcat-3/

2 with a more automated process taking advantage of the standard protocols and schemas offered by Zenodo. This task involves a number of challenges that we are currently addressing with regard to the selection of the source API, the selection and conversion of metadata, the selection of a subset of the downloaded metadata records and the setting-up of an automated procedure for regular harvesting.

5 Language Resources in the ELG Catalogue

After the most recent ingestions of datasets as well as the contributions from the pilot projects and ELE, the ELG catalogue has reached a total of 8,873 metadata entries in April 2022, far exceeding our expectations when we started the project. The majority of these are description records without the data being hosted in ELG (103 resources are fully available through ELG). However, even if not available through ELG directly, most datasets are available through the referenced repository page, often available for download, which is reflected in the ELG catalogue too. Figures 1 and 2 illustrate the breakdown of repository sources ingested so far together with the breakdown of the current numbers per source.

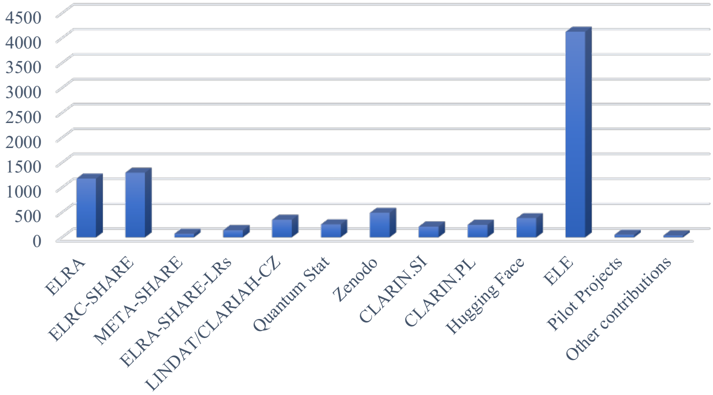

Fig. 1 Repository sources of the 8,873 datasets available in ELG in April 2022

Regarding resource types and their linguality, Figure 3 illustrates the numbers. As expected, the highest numbers apply to corpora (6,236 available in ELG), with twice as many monolingual corpora as bilingual ones (which in turn are three times as many as the multilingual ones). Lexical/Conceptual resources are also very well represented with 2,229 entries.

One of our bigger concerns at the time of Release 2 was the fact that there were barely any language descriptions (there were only 7). This has changed with the work towards ELG Release 3: at the time of writing, we count 408 language descriptions with the majority being monolingual. Further regarding language descriptions, the

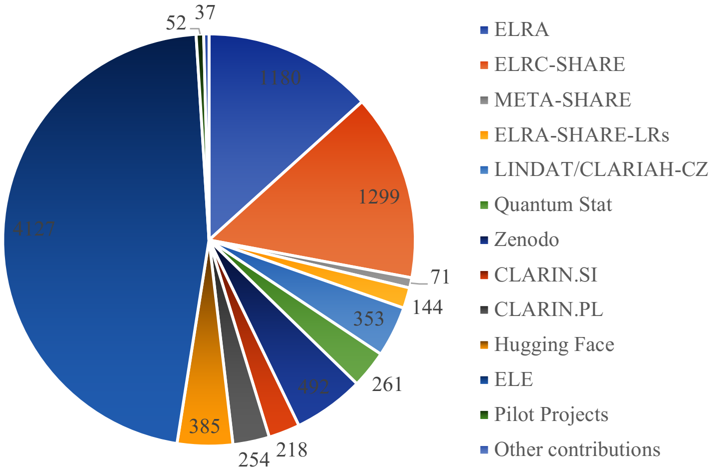

Fig. 2 Repository sources of the 8,873 datasets available in ELG in April 2022

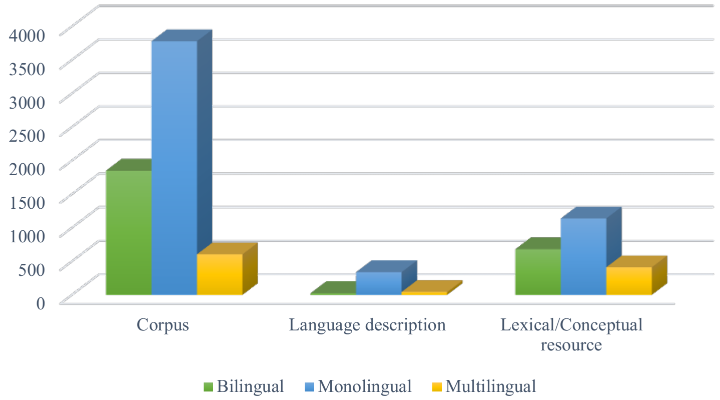

Fig. 3 Types of resources according to linguality

number of its "language models" subclass has increased to 358. This is good news as models are a popular and highly demanded resource type, currently providing the state of the art for many LT/NLP tasks. ELG is actively encouraging the use of its platform for the creation of models. The pilot projects have supported this resource type as well by contributing their models, too.

ELG also offers a very broad language coverage, with 450 languages represented by lexical/conceptual resources, and with corpora available in 438 languages, at the time of writing. The language models cover 156 languages, grammars are available for 25 languages. These are either monolingual or multilingual resources. Figure 4 shows the language resource type distribution for the EU official languages.

Finally, different media types are also represented in ELG. As expected, the largest number of resources belongs to the type "text" with more than 7,000 datasets.

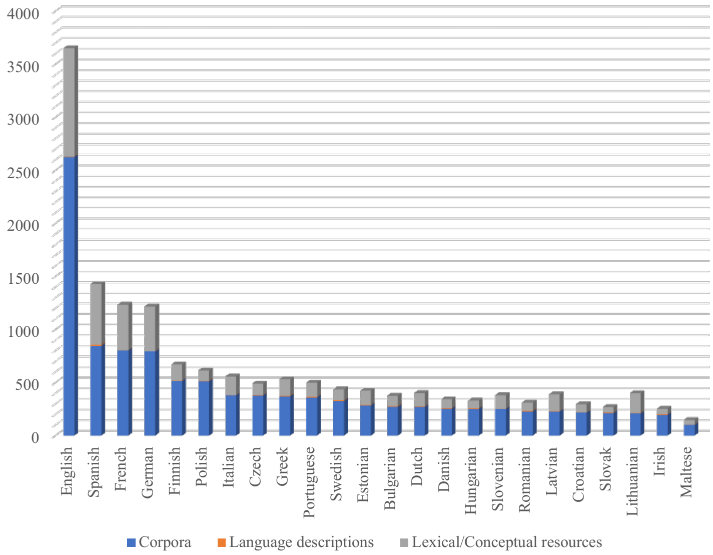

Fig. 4 Language resource type distribution for the official EU languages

Nonetheless, the type "audio" already offers more than 1,200 resources while currently 385 image and video resources are available.

6 Language Resources and Legal Issues

Managing legal issues in a large platform such as ELG implies taking care of a wide variety of legal aspects, often regarding licensing. It also implies taking into account processes that may differ from one provider to another. A provider may choose to distribute resources either through implicit or explicit licences, through specific conditions of use, or through considering a particular user status such as profiles or membership status. Moreover, the need to ensure GDPR compliance requires certain monitoring processes. For the development of the platform, the project has benefited from the support and advice of a dedicated team of legal experts who helped deploy the platform in a manner that is legally sound. This ranges from establishing the necessary legal context (e. g., Privacy Policy and Terms of Use) to stepping in for consultations. The legal team has also contributed to the preparation of a Data Management Plan (see Section 7). Below, we briefly describe some of the specific issues the ELG legal team has taken care of.

Advice on implicit versus explicit licences: One main distinction to make is the management of implied (or implicit) versus expressed (or explicit) licences. For implied licences, it has become a commonly and widely used practice to grant users access when they click on the licence terms acceptance button indicated on the repository pages.

Advice on conditions and terms of use: The conditions of use of a resource are another factor that has been defined 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 need to review the purpose of use (which could be commercial, for research, etc.), as well as the profile of the licencee (this is the type of institution, some resources may be restricted to particular types of institutions, e. g., *academic* or *commercial*)[26].

Financial and distributional 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. Parameters like the legal profile of the licencee, the purpose of use and the pricing policy need to be clearly displayed.

META-SHARE licensing: The selection of LRs for ingestion done for the three META-SHARE nodes needed to be revised due to licensing restrictions. These involved proprietary licences (e. g., MS-C-NoReD, MS-NC-NoReD and MS-Commons-BY-SA), as well as licences that required negotiations with providers. To address this, a study of the licences was performed by the ELG legal team for discussion with node managers. A proposal for licence mapping was drafted where non-restrictive licences were invited to move to Creative Commons licences. Restrictive licences were encouraged to move to more open licences, too.

Legal checking: The identification of various repositories demonstrated the importance of legal checking all throughout the information compilation process. In some cases (e. g., Zenodo), licences were well identified and could usually be integrated in the ELG metadata without further analysis. However, for other cases (e. g., ELRA-SHARE-LRs, Quantum Stat), legal information did not always comply with ELG requirements or was simply missing. Consequently, legal expertise was needed to either check and confirm the accuracy of present legal information, or to search for and gather the appropriate legal information.

Improvement of the licence list: When we processed the Zenodo datasets, we realised that several licences were not part of the ELG metadata values. Thus, the ELG legal expert was asked to compare the Zenodo list with the ELG list and make suggestions to integrate some of those licences into the ELG metadata. A list of 68 licences that did not correspond to ELG values was checked, out of which 40 could be added to the ELG licence list, whereas the other 28 did not need to be added because they were already used within ELG using other labels, they were not used, or they had no link.

Addition of conditions of use in the ELG metadata: We decided to add a new metadata field corresponding to the “conditions of use” associated to each identified licence to improve the search functionality for resources based on their licensing conditions. For “standard” licences, the conditions of use were added by the ELG team, based on information gathered from Creative Commons licences,

[26] https://live.european-language-grid.eu/terms-of-use

values from the CLARIN licencing framework[27], META-SHARE licences, and the ELRA licence wizard[28]. For all other LRs, a thorough analysis of over 300 licences (all licences in the SPDX list[29]) was done by our legal team who went through the different conditions of use such as the intellectual property rights granted by the licences, the requirements on redistribution imposed by the licence, the requirements on use of the data and, finally, the requirements imposed on users (Rigault et al. 2022b).

7 Language Resources and Data Management

ELG is a platform for commercial and non-commercial Language Technologies, both functional (running services and tools) and non-functional (datasets, resources, models). In order to achieve this, the consortium in charge of the ELG platform has enacted several priorities that include the processing of massive amounts of data and of different types. These large amounts of data derive from partners' contributions, external providers willing to share their datasets through ELG, our harvesting of other repositories as well as different kinds of resource and repository identification work. As can be expected, such a data intensive project requires clear data management policies, in particular considering GDPR constraints. For that purpose, we implemented a Data Management Plan (DMP) as a concrete necessity for organisational, technical and legal management of all data types processed in the course of the project (Rigault et al. 2022a). The DMP documents the variety of data types collected, received and/or processed in the course of the project and reports on how the data is going to be managed with regard to technical, organisational and legal aspects. The DMP also complies with best practices and, in particular, with the requirements of Horizon 2020 as well as GDPR obligations. It defines useful practices to enhance compatibility with the FAIR principles (see Section 7 in Chapter 2 and Wilkinson et al. 2016)[30], as endorsed and specified for Horizon 2020. Moreover, the DMP provides advice in terms of best practices for language resource creation in all steps of an LR life cycle (Choukri and Arranz 2012; Rehm 2016).

8 Conclusions

We integrated more than 10,000 metadata records for datasets, models and other classes of language resources into the ELG platform. These LRTs have been carefully described so as to ease their findability (following the FAIR principles) and to

[27] See https://www.clarin.eu/content/licenses-and-clarin-categories#res and https://www.clarin.eu/content/clarin-license-category-calculator

[28] http://wizard.elra.info/principal.php

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

[30] https://www.go-fair.org

ensure compliance with the ELG metadata schema while advocating for interoperability. A series of steps and best practices has been followed with the objective of establishing procedures for resource identification, description and ingestion. The work carried out during the ELG project has allowed us to consider expertise and lessons learned to improve protocols and principles. This has been the reason for updating the integration approach of some repositories (e. g., ELRC-SHARE and Zenodo). The strategy behind the choice of repositories has also been planned carefully, following technical and strategic priorities, as well as evolutionary needs and demands. ELG users can now either access thousands of resources or contribute resources through the different means provided. Legal issues have also been considered with a special focus on licensing. Moreover, a Data Management Plan has been conceived to address the handling of all types of data (including sensitive data) within ELG as well as guiding the production and life cycle aspects of LRs.

References

Arranz, Victoria, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jan Hajic, Ondrej Kosarko, Cristian Berrio, Andrés Garcia-Silva, Rémi Calizzano, Nils Feldhus, Miltos Deligiannis, Penny Labropoulou, Stelios Piperidis, and Ulrich Germann (2021). *Deliverable D5.2 Data Sets, Identified Gaps, Produced Resources and Models (Version 2)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.2-final.pdf.

Arranz, Victoria, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Penny Labropoulou, Miltos Deligiannis, Leon Voukoutis, Stelios Piperidis, and Ulrich Germann (2022). *Deliverable D5.3 Data Sets, Models, Identified Gaps, Produced Resources and their Exploitation within ELG (Version 3)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.3-final.pdf.

Choukri, Khalid and Victoria Arranz (2012). "An Analytical Model of Language Resource Sustainability". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 1395–1402. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/846_Paper.pdf.

Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli (2012). "The META-SHARE Metadata Schema for the Description of Language Resources". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 1090–1097. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3421–3430. URL: https://www.aclweb.org/anthology/2020.lrec-1.420/.

Lösch, Andrea, Valérie Mapelli, Khalid Choukri, Maria Giagkou, Stelios Piperidis, Prokopis Prokopidis, Vassilis Papavassiliou, Miltos Deligiannis, Aivars Berzins, Andrejs Vasiljevs, Eileen Schnur, Thierry Declerck, and Josef van Genabith (2021). "Collection and Curation of Language Data within the European Language Resource Coordination (ELRC)". In: *Proceedings of the Conference on Digital Curation Technologies (QURATOR 2021)*. Ed. by Adrian Paschke, Georg Rehm, Jamal Al Qundus, Clemens Neudecker, and Lydia Pintscher. Vol. 2836. CEUR Workshop Proceedings. Berlin, Germany: CEUR-WS.org. URL: http://ceur-ws.org/Vol-2836/qurator2021_paper_6.pdf.

Mapelli, Valérie, Victoria Arranz, Hélène Mazo, and Khalid Choukri (2022). "Language Resources to Support Language Diversity – the ELRA Achievements". In: *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 551–558. URL: http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.58.pdf.

Piperidis, Stelios, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi (2014). "META-SHARE: One year after". In: *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: ELRA, pp. 1532–1538. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/786_Paper.pdf.

Rehm, Georg (2016). "The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources". In: *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: ELRA, pp. 2450–2454. URL: https://aclanthology.org/L16-1388.pdf.

Rehm, Georg and Katrin Marheinecke (2019). *Deliverable D7.2 National Competence Centres and Language Technology Council*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2021/02/ELG-Deliverable-D7.2-final.pdf.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rigault, Mickaël, Victoria Arranz, Khalid Choukri, Valérie Mapelli, Pawel Kamocki, and Lucille Blanchard (2022a). *Deliverable D5.6 Data Management Plan (Version 3)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.6-final.pdf.

Rigault, Mickaël, Victoria Arranz, Valérie Mapelli, Penny Labropoulou, and Stelios Piperidis (2022b). "Categorizing Legal Features in a Metadata-Oriented Task: Defining the Conditions of Use". In: *Proceedings of the Legal and Ethical Issues Workshop (LREC 2022)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 22–26.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18. URL: http://www.nature.com/articles/sdata201618.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2020). "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Yeganova, Lana, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de-Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes (2021). "Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set". In: *Proceedings of the Sixth Conference on Machine Translation*. ACL, pp. 664–683. URL: https://aclanthology.org/2021.wmt-1.70.