EUROPEAN LANGUAGE GRID

D5.3

Data sets, models, identified gaps, produced resources and their exploitation within ELG (version 3)

Authors:	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Mickaël Rigault (ELDA); Penny Labropoulou, Miltos Deligiannis, Leon Voukoutis, Stelios Piperidis (ILSP); Ulrich Germann (UEDIN)

Dissemination Level:	Public
Date:	31-01-2022

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 42 months (GA amendment version: AMD-825627-7)
Deliverable number	D5.3
Deliverable title	Data sets, models, identified gaps, produced resources and their exploitation within ELG (version 3)
Type	Report
Number of pages	27
Status and version	Final - Version 1.0
Dissemination level	Public
Date of delivery	Contractual: 31-01-2022 – Actual: 31-01-2022
WP number and title	WP5: Grid Content: Language Resources, Datasets, and Models
Task number and title	Task 5.1: Identification and collection of existing data sets and resources to make them available through the ELG; Task 5.2: Identification of severe LR gaps and creation of LRs; Task 5.3: Dynamic training of new models in the ELG; Task 5.4: Legal support, DMP and GDPR
Authors	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Mickaël Rigault (ELDA); Penny Labropoulou, Miltos Deligiannis, Leon Voukoutis, Stelios Piperidis (ILSP); Ulrich Germann (UEDIN)
Reviewers	Katrin Marheinecke (DFKI); Katja Prinz (HENS)
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
	Hensoldt Analytics GmbH (HENS), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Miklos Druskoczi
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Prof. 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

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CC	Creative Commons
COAR	Controlled Vocabularies for Repositories
CSV	Comma-separated Values
DC	Dublin Core
DCAT	Data Catalog Vocabulary
DOC	Microsoft Word File (file extension)
EC	European Commision
ELE	European Language Equality
ELG	European Language Grid
EOSC	European Open Science Cloud
EU	European Union
HF	Hugging Face
HLT	Human Language Technology
JSON	JavaScript Object Notation
LR or LRs	Language Resource or Language Resources
LRTs	Language Resources and Tools
LT or LTs	Language Technology or Language Technologies
ML	Machine Learning
R2	Release 2
R3	Release 3
RDF	Resource Description Framework
REST	Representational state transfer
SKOS	Simple Knowledge Organisation System
SPDX	Software Package Data Exchange
TMX	Translation Memory Exchange
TRL	Technology Readiness Level

WAV	Waveform Audio File Format
WMT	Workshop on Machine Translation
XML	Extensible Markup Language
XSD	XML Schema Definition

Abstract

This report describes the work carried out to populate the European Language Grid since D5.2 (M25), taking into account that Release 2 (R2) took place in M26 (February 2021) and that we are close to the pre-final Release 3 (R3) in M38 (February 2022). This deliverable will report on a) remaining datasets from repositories under integration in the ELG for R2; b) new repository priorities for Release 3; c) new metadata conversion, harvesting and description work performed since R2 to manage both ongoing and new repositories; d) the technical and legal issues addressed for R3, which includes solving issues as well as looking at improvements; e) status of the ELG after analysis of its content and of the existing gaps. The report also provides an overview of the experience in using ELG data for Model Training that had been described in D5.2.[1] Finally, we highlight where we stand and what next steps are foreseen for the final months of the project.

1 Introduction

D5.3 "Data sets, models, identified gaps, produced resources and their exploitation within ELG (version 3)" describes the work carried out in the past 12 months to populate the European Language Grid catalogue. D5.3 follows on the reporting done for D5.2 by:

- Explaining the evolution of already initiated repositories in January 2021 (at the time of R2).
- Presenting new repository priorities for Release 3.
- Describing the different approaches for dataset integration, which vary from metadata conversion to harvesting, and which address metadata description and validation.
- Furthermore, D5.3 also details all the technical and legal issues that have been addressed for R3, where the former have a strong legal context, trying to lighten up the conditions of use as well as the ELG metadata schema as a whole. The legal issues will also introduce the analysis of ELG and SPDX licences[2] that has been performed. Once all ingestion repositories have been addressed and other contributions explained, we will look into the current content of the ELG catalogue, with statistics and considerations of what we offer to the user so far and with some thoughts of how reported gaps are being handled. As a pre-concluding section, we offer a sort of return on experience after using the platform for model training. This experience derives from Task 5.3, which was concluded by D5.2 and which has allowed for some initial testing and thoughts on lessons learnt. These sections are followed by some conclusions drawn on the work presented.

2 Provision of Language Resources for ELG Release 3

Each ELG release has followed an evolutive strategy for catalogue population. This strategy has evolved as procedures have been put into place and new priorities and needs have been defined. Release 2 launched an am-

[1] ELG Deliverable: D5.2 – Data sets, identified gaps, produced resources and models (version 2) (January 2021).
[2] https://spdx.org/licenses/

bitious acquisition of very large catalogues which were not compliant with ELG's structure and metadata description. This was the case, for instance, for Quantum Stat and Zenodo[3]. Repositories like Zenodo are extremely large digital libraries where many different research artefacts are published and, thus, the user needs to do a somehow archaeological work to extract relevant content. Despite these difficulties, the outcome is rewarding as it provides access to many Human Language Technology (HLT) related datasets, a diverse and rich content made available to the community.

The Language Resource (LR) provision for R3 has built on the processes and strategies established in the two earlier releases, expanding its objectives as follows:

1. Continue and conclude (if possible) with already initiated repositories for R1 and R2 (we have adopted harvesting-based updating procedures for some repositories and, thus, these are not concluded, but updated regularly as they are populated themselves[4].
2. Conclude conversion and description procedures for both Quantum Stat and Zenodo.
3. Plan for the establishment of automatised protocols to address Zenodo's updates.
4. Move ahead with remaining nodes from the META-SHARE network.
5. Define repository priorities for R3.
6. Set up harvesting procedures for as many ingested repositories as possible to automatise means and guarantee updates.

2.1 Remaining Resources and Updates from already initiated Release-2 Repositories

Following the objectives mentioned above, during this period we have concluded conversion and description procedures for some repositories that were under processing for R2 (ELRA-SHARE-LRs 2014, 2016, 2018 and 2020[5], Quantum Stat and Zenodo). In particular Quantum Stat and Zenodo have required large efforts in terms of conversion, mapping and description. Their metadata is very limited and already the filtering efforts to locate and reuse the relevant entries is very costly. This has led us to engage in more sophisticated automatised protocols, for instance, for Zenodo, as it will be seen in Section 3.4.

Furthermore, we have continued the regular harvesting of the two repositories initiated in previous releases, namely LINDAT/CLARIAH-CZ[6] and ELRC-SHARE[7]. Apart from minor updates that have been made by both teams to conform with the updated ELG releases, the harvesting has continued without problems.

Discussions have also advanced towards the integration of further META-SHARE nodes, even if the approach followed has been adjusted to the current situation, new priorities, and potential new approaches. This is further explained in Section 3.3.

The work carried out to ingest all these repositories is detailed in Section 3 and Table 1 shows the figures achieved (including also those ingested for R1), adding up to 3,800 datasets in this first step.

[3] As explained in D5.2.
[4] For example, ELRC-SHARE, LINDAT/CLARIAH-CZ, among others.
[5] These resources have been reported as LREC Shared LRs in D5.2 and have been renamed as ELRA-SHARE-LRs.
[6] https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en
[7] https://www.elrc-share.eu

	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	–	1180
ELRC-SHARE	1249	50	–	1299
META-SHARE	52	12	7	71
ELRA-SHARE-LRs	105	37	2	144
LINDAT/CLARIAH-CZ	274	79	–	353
Quantum Stat	255	6	–	261
Zenodo	328	129	35	492

Table 1: Remaining LRs and updates from initiated R2 repositories

2.2 Repository Priorities for ELG Release 3

For Release 3 (M38), we are looking into repositories and catalogues that include LRTs of potential interest to ELG consumers and that could be harvested automatically. Following the objectives defined in Section 2, priorities for the ingestion of new repositories were set based on the following:

- Relevance of the resources for Language Technology (LT) purposes.
- The size of the catalogue contents, and
- The availability of an export mechanism for the exchange of metadata, preferably according to a standard harvesting protocol.

The selected sources for this stage in ELG population have been specific national CLARIN centres (CLARIN.Si and CLARIN.PL), Hugging Face, and Zenodo (Phase 2)[8]. Section 3 reports on the processes undertaken for the conversion of their contents and import into ELG, as well as on the outcomes.

The harvesting work performed on the two CLARIN centres and on Hugging Face has resulted on the ingestion of the further 857 datasets that are detailed in Table 2.

	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
CLARIN.SI	140	78	–	218
CLARIN.PL	239	15	–	254
Hugging Face	385	–	–	385

Table 2: LRs ingested from R3 repository priorities through harvesting

The harvesting procedure for Zenodo is currently being set up as it is described in Section 3.4, so no further resources have been ingested from this repository in addition to the phase 1 approach referred to in Section 2.1. However, harvesting is expected to be launched soon, a large number of datasets is foreseen to be obtained[9].

One additional data source has been defined as priority for ELG and is under analysis for inclusion:

[8] A second phase has been planned to integrate Zenodo updates through a regular harvesting protocol.
[9] 592,509 metadata records have been obtained from Zenodo (cf. Section 3.4.2) using the "dataset" and "software" filters. These will be filtered down further when selecting the relevant entries.

- Data from the different WMT (Workshop on Machine Translation)[10] events that have been held since 2006. Data stored after the different workshops, evaluation tasks and other shared tasks present a very idiosyncratic structure which prevents us from deploying already established protocols in a straightforward manner. Details on the status of this work are provided in Section 3.7.

Finally, in addition to the ingested datasets that have been listed in both Table 1 and Table 2, Section 5.2 is also dedicated to the procedures and outcomes of a crowdsourcing survey for LRTs conducted in the framework of the European Language Equality (ELE)[11] project, the results of which have been included in the ELG platform under the umbrella of a close collaboration between both projects. This has contributed to the enriching of the ELG platform with 4,127 datasets.

At present, the ELG catalogue hosts 8873 datasets and further population is going to take place before the end of the project in June 2022. Figure 1 and Figure 2 illustrate the summary of the sources ingested so far together with the breakdown of the current numbers per source.

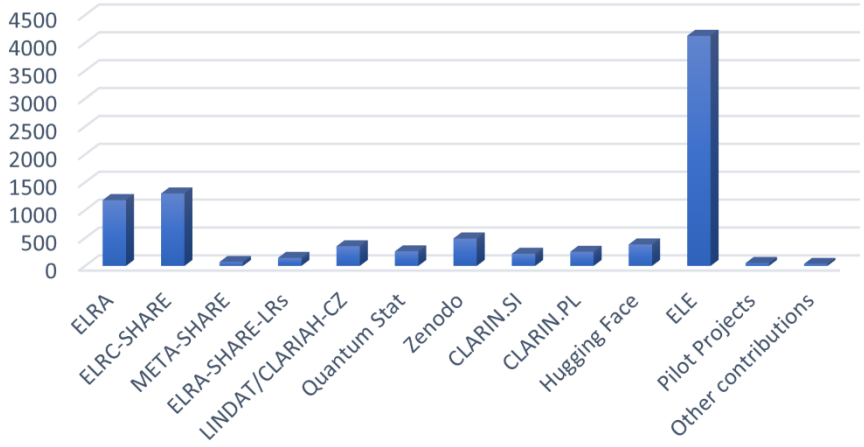

Figure 1: Breakdown of the 8873 datasets in ELG and their respective sources

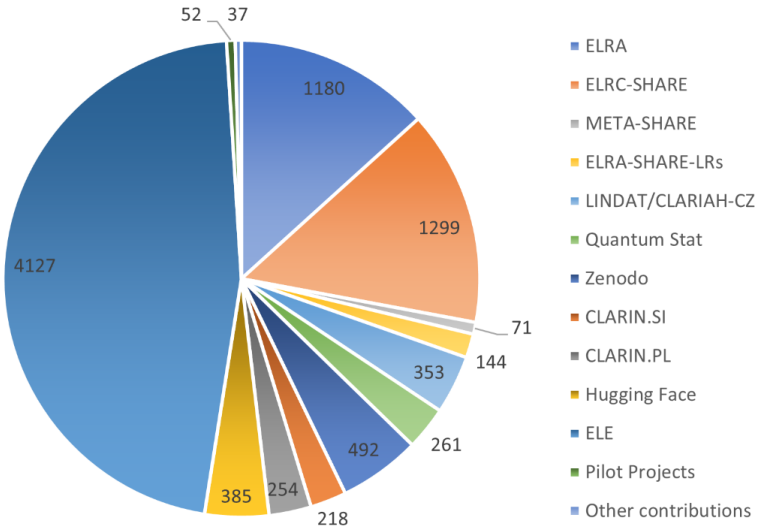

Figure 2: Breakdown of the 8873 datasets in ELG and their respective sources

[10] https://www.statmt.org/wmt21/index.html
[11] https://european-language-equality.eu

3 Language Resource Metadata Conversion, Harvesting and Description

This section describes the procedures to populate the ELG catalogue with the content from the different priority source repositories listed in Sections 2.1 and 2.2. Those repositories for which processing had already been initiated will describe their follow-up steps, namely, implemented improvements or adjustments. Datasets deriving from collaborations such as ELE and the ELG-financed pilot projects are described in Section 5, which is dedicated to content population following the analysis and identification of gaps under a joint strategy.

3.1 ELRA-SHARE-LRs

As a reminder to their introduction in D5.2, the ELRA-SHARE-LRs are provided by conference participants attending LREC, the Language Resources and Evaluation Conference[12]. This initiative was launched in 2014 and the contributed language resources have been made available through the ELG catalogue. R2 counted on those from the 2014, 2016 and 2018 events, and Release 3 has ingested those that were shared at the LREC 2020 conference. Following the publication of the ELRA-SHARE-LRs 2020[13] online, we proceeded with their ingestion into the ELG platform in March 2021. As for the previous packages (2014, 2016 and 2018), a selection of LRs was done by checking the compliance of licences with the ones accepted in ELG and by completing the excel template file. Then, the excel file was converted into a series of XML files that could be ingested into the ELG platform through the editor Dashboard.[14]

The ELRA-SHARE-LRs 2020 package published 73 additional language resources, resulting in the total 144 reported in Table 1.

3.2 Quantum Stat

As established in D5.2, the work initiated for Release 1 has been finalised towards Release 2. Out of the list of 481 LRs exported initially, our revision and filtering led us to restrict the final list to a smaller set of 261 LRs (see Table 1). As already explained in different occasions, datasets contained in Quantum Stat are poorly described and are often irrelevant for HLT. Thus, revision of all exported dataset candidates was a crucial step of their preparation. We proceeded with the ingestion of the filtered, described and converted datasets into the ELG platform in February 2021.

3.3 META-SHARE

The META-SHARE network was analysed to define the optimal strategy for its inclusion in the ELG platform (see D5.2). The three nodes already ingested so far (META-SHARE-DFKI, META-SHARE-ELDA and META-SHARE-ILSP) followed an approach of metadata conversion through the implementation of metadata schema converters[15]. These nodes are managed by ELG project partners, and the procedure was straightforward. However, the next steps proved to be more complex than expected, due to several considerations:

- The maintenance behind the different nodes differs, which implies that nodes have followed different paths in their evolution and while some may have moved forward very dynamically in their LRT sharing objectives towards other ventures (e.g., CLARIN infrastructure), some others may have remained more static and even obsolete.

[12] https://lrec2022.lrec-conf.org/en/
[13] https://lrec2020.lrec-conf.org/en/shared-lrs/
[14] For further details on the selection and metadata description, please refer to ELG Deliverable D5.2.
[15] ELG Deliverable "D5.1: Identification and collection of existing datasets, models, identified gaps and plans (version 1) (April 2020).

- This lack of homogenisation in the nodes content has an impact on both the sharing strategy (do managers wish to have their node integrated in an infrastructure such as ELG?) as well as with regard to the manner to address the description metadata that needs to be converted/mapped as it has evolved differently depending on the node.

This has made us change the strategy ourselves and redesign the way the integration of the META-SHARE network may be carried out:

- By looking at the nodes as members of application groups, where those connected to other models may be addressed with a group model in mind.
- Group models allow us to design a single ingestion protocol (even if some minor node-specific tuning is required) for several nodes. Such a group-model approach has been followed, for instance, for the three CLARIN nodes which are already in ELG (see Section 3.5).
- By being more flexible about the migration of the datasets themselves into ELG, as META-SHARE nodes which have been converted or connected to CLARIN, for instance, guarantee a certain maintenance and dataset storage, and can just export their metadata if they wish so.

At present, the META-SHARE-ILC node[16] is under discussion with its managers given that it is a good example of a node that has migrated its data into the CLARIN infrastructure (as CLARIN ILC-CNR node) and that may be ingested into ELG through this new access rather than through its older META-SHARE access, given that it may allow revisiting the harvesting procedures that have been put into place for other CLARIN nodes.

3.4 Zenodo

As introduced in D5.2 and in Section 2.1 of the current document, the original process behind the insertion of the Zenodo repository into ELG proved its limitations due to:

- The extremely time-consuming effort needed to perform the extraction, conversion, and description of the repository's content, as well as
- The small number of LRs we managed to obtain compared to the high number of entries in the Zenodo Library. A long filtering task needs to be envisaged with every database export/extraction.

Thus, we moved towards a more automatised process to address the large number of metadata entries that we were targeting to extract. This is explained in the following section.

3.4.1 Semi-automatic means to ingest Zenodo records

Zenodo offers a Rest API that allows harvesting metadata, available under Creative Commons CC-0 licence. To obtain more LRs for ELG, we decided to query the Zenodo database directly using a script.

With the REST API, we broadened the File Type criteria (zip, html, xml, doc, docx, txt, xlsx and csv) and searched for resources using the following keywords:

- lexicon
- lexica
- corpus

[16] http://metashare.ilc.cnr.it

- corpora
- lexical database
- terminology
- terminologies
- terminological
- glossary
- glossaries

This produced a list of 719 new resources in JSON format.

The compilation of metadata information still required manual operation to make the selection of the actual LRs to be ingested as well as add the minimal set of metadata elements which are mandatory for ELG and which do not exist in the Zenodo records:

- To ease the conversion, the Excel template was adapted in order to have one single file to cover the different types of LRs (Written corpora, Lexical Conceptual Resources, Audio Corpora, Video Corpora) in one single folder.
- Datasets that were not considered as Language Resources were discarded.
- Other types of LRs not fully covered by the Excel template and conversion tool (language descriptions, multimodal LRs and image corpora) had to be treated separately and manually through the ELG Editor.
- A Python script was developed to convert the remaining selected resources into Excel to facilitate the manual entering of the missing metadata.

As of the writing of this document, we have managed to submit 492 LRs into the ELG platform out of the 719 originally extracted. 122 datasets were discarded as they were not considered as Language Resources. A remaining set of 105 LRs are being checked for Release 3 at the moment. The conclusion reached for this semi-automatised process was that it was still very costly and could not be executed in an agile regular manner to integrate updates. Therefore, an automatic harvesting-oriented approach is currently being designed and is described in the following subsection.

3.4.2 Harvesting Zenodo

The constant update of the Zenodo catalogue[17] and its uptake by researchers for the upload of datasets, and, most recently, software, makes it particularly interesting for ELG purposes. Zenodo exposes the metadata records in two channels: through a REST API[18], which outputs records as JSON files, and an OAI-PMH API[19] in a set of standard metadata formats, namely DC[20], DataCite[21], MARC21[22] and DCAT[23]. Work is currently ongoing to replace the semi-manual import of metadata records of Zenodo that started in the previous release with a more automated process taking advantage of the standard protocols and schemas offered by Zenodo.

This task involves a number of challenges we are currently addressing:

[17] https://zenodo.org
[18] https://developers.zenodo.org/#rest-api
[19] https://developers.zenodo.org/#oai-pmh
[20] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[21] https://schema.datacite.org/meta/kernel-4.4/
[22] https://www.loc.gov/marc/bibliographic/
[23] https://www.w3.org/TR/vocab-dcat-3/

- **Selection of the source API**: The preferred API for the ELG import is the OAI-PMH protocol, which is a standard protocol for interoperability and exchange of metadata records and includes a mechanism for regular harvesting. However, this endpoint doesn't offer a query set for “datasets” and “software” (which are the two resource types of interest to ELG), a choice that is available through the REST API; in addition, harvesting from the OAI endpoint for big amounts of metadata records is unreliable. Thus, we have resorted to a dual solution:
    - We have downloaded the automatically generated full dump of 2,060,674 metadata records that were included in Zenodo until 31/08/2021; this dump contains all the records in JSON format and is available from Zenodo.
    - For records added to Zenodo after this date, we are doing an incremental harvesting from the OAI-PMH endpoint. Through this channel, we have downloaded 147,621 more records during a three-month period.
- **Selection and conversion of metadata**: For the reasons described above, we are implementing two converters:
    - Mapping from the JSON[24] element offered through the REST API to the ELG schema is ongoing.
    - With regard to metadata formats offered through the OAI-PMH endpoint, we opted for the DCAT metadata schema, which is currently one of the most popular schemas across repositories. Thus, we expect that mapping DCAT into ELG will allow us to re-use the same converters as a base for import from other repositories. Moreover, DCAT is the schema with the richer information among the ones exposed from Zenodo, and the only one that includes a direct link to the downloadable files (*downloadURL* element), an important feature for ELG consumers. Our aim is to use an XSLT converter for this endeavour, given that the OAI-PMH endpoint makes available the records in an XML format. However, the mapping from DCAT is not straightforward; DCAT is an RDF vocabulary and restrictions, and extensions are implemented in the form of profiles and applications. The application profile or the XSD schema[25] used by Zenodo is not publicly available. In addition, a closer inspection into the XML files has revealed discrepancies in the representation of the same element. For instance, subject (which is defined in DCAT as a SKOS Concept) appears in Zenodo XML files either as a SKOS[26] Concept or as an element with the IRI of the subject value in the form of an attribute. We have analysed the Zenodo XML files to the extent possible and we are now in the mapping phase.
    - **Selection of a subset of the downloaded metadata records**: From the 2,208,295 metadata records available until 31/12/2021 those with resource type “dataset” and “software” amount to 592,509 entries. Such a number is obviously too high, and the majority of these records is of little or no interest at all to ELG users[27]. We are experimenting with various filters from the Zenodo metadata elements records and statistical measures to identify candidate records for our import. The filters that we will select will be used for the automatic identification of records to be imported in ELG when the harvesting is put in place.

[24] https://developers.zenodo.org/#representation
[25] The XSD schema included in the OAI-PMH API for the DCAT schema is in fact that of DataCite v4.1.
[26] https://www.w3.org/2004/02/skos/
[27] As a comparison, the ELG catalogue, following the import of metadata from the crowdsourcing initiative launched by ELE, currently has around 11,000 metadata records.

- **Setup of an automated procedure for regular harvesting**: This process is still under discussion, until we have the first results from our experiments and the converters implemented.

3.5 CLARIN-SI and CLARIN-PL

The LINDAT/CLARIAH-CZ repository makes available an OAI-PMH endpoint which exposes ELG-compatible metadata records. The same repository software (developed by the LINDAT team based on DSpace) is used by several other CLARIN centres for the setup of their repositories. This makes them ready-to-import into ELG using the same harvesting mechanism and procedure.

For this release, this collaboration has resulted in the regular harvesting of two more CLARIN centres: the **Slovenian CLARIN** (CLARIN-SI)[28] and the **Polish CLARIN** (CLARIN-PL)[29]. Discussions are ongoing with one of the Italian CLARIN nodes (CLARIN-IT)[30], which derives partially from the Italian META-SHARE-ILC node (Section 3.3).

Moreover, there is ongoing work for the extension of the harvested resource types. Currently, the harvesting is restricted to corpora and lexical/conceptual resources, and the intention is to import also the remaining types, i.e., models, grammars and tools/services. For this extension, the ELG and LINDAT teams are collaborating on the required mappers and converters.

3.6 Hugging Face

The catalogue of Hugging Face[31] includes a large collection of ML (Machine Learning) models and datasets that can be used for training models, with a focus on transformers. ELG has entered into a collaboration with Hugging Face for importing metadata records from their catalogue.

The primary aim of Hugging Face is to enable its users to upload datasets and models following a set of specifications so that they can be deployed for testing and building other models and/or integrating models in their applications. Although they encourage adding descriptions for the resources, this is not strictly enforced, and the suggested metadata elements do not follow a standard schema. Users are asked to upload a "card" for datasets[32] and models[33], with a combination of free text fields and a set of tags (e.g., language, licence) with values from recommended controlled vocabularies (which are, however, not strictly validated). The emphasis of the import into Hugging Face is on the upload of the content files rather than their description.

Hugging Face exposes two distinct APIs with JSON files for datasets and models respectively. These JSON files include a subset of the metadata elements displayed on their catalogue. For datasets, these are: id (which serves also as the resource name), description, citation, author(s), and a set of tags for language, multilinguality, size category, licence, task category and task identifier (subtask). For models, the only elements are the id (serving as the name) and the task category of the model. In addition, not all records have values for all of the above elements, given that the original providers have not filled them in. To be imported into ELG, however, the metadata records must comply with the ELG metadata schema, which means that at least the mandatory elements of the minimal version[34] are filled in. For this reason, the conversion and import of records from Hug-

[28] http://www.clarin.si/info/about/
[29] https://clarin-pl.eu/dspace/
[30] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/
[31] https://huggingface.co
[32] https://huggingface.co/docs/datasets/dataset_card.html
[33] https://huggingface.co/docs/hub/model-repos
[34] https://european-language-grid.readthedocs.io/en/stable/all/A2_Metadata/MinimalVersion.html

ging Face into ELG was restricted to datasets and only to those that have filled in at least a description, and values for the language and licence elements, which are deemed the minimum threshold for the findability and usability purposes in the context of ELG.

A custom converter was developed based on the mapping of the elements and, in the case of controlled vocabularies, their values. Manual work was further necessitated for the enrichment of information for specific elements. The most prominent case was that of licences: ELG requires, besides the name of the licence, a URL with the text of the licence. Hugging Face includes a list of licence identifiers taken from the SPDX list (which are also used in ELG), but also allows users to add a licence name without further information. Thus, in addition to the mappings of the licence identifiers from Hugging Face into the ones used in ELG, we looked for the licence URL of unmapped values; if no URL was found, the resource was not imported into ELG. Finally, where required, default values have been used for mandatory elements whose values could not be inferred from the original metadata records (e.g., all datasets have been assigned the text value for media type). Annex A describes the mapping of the elements.

3.7 WMT

As introduced in Section 2.2, all the data generated in the context of the different WMT events is very relevant to ELG. Besides the data itself, interest also lies on the service that will be provided to future users that may consider access to this data extremely complicated. Datasets for different evaluation and shared tasks are spread in between different pages in a complex manner which does not allow for a quick and simple usage. Moreover, data documentation and structure vary from event to event, which increases the difficulty. Last but not least, some datasets are shared across evaluation tasks, which implies a risk of data duplication.

Following a preliminary analysis of the source(s), we are currently studying WMT's data structure to see how to extract datasets as well as how to represent them and offer them to the users in a clear re-structured and well documented manner. For that purpose, the data structure for some priority years is being looked into and mapped into a content tree for later mapping into the ELG platform.

Work on this will be reported at the end of the project.

4 Technical and Legal Issues Addressed for Release 3

4.1 Technical Issues

Technical issues that are specific to each source have been reported in the respective sections. In this section, we describe the changes that were introduced in the ELG schema (version 3.0.0) and platform. More specifically, a relaxed version of the schema has been created which allows us to import metadata records from other catalogues with poorer information or documented with more general schemas. These changes are allowed mainly for metadata records marked as "for information"[35]:

- Changed the optionality status for specific elements.

[35] It is currently being discussed, whether a different term for these resources would be more useful for users because "for information" might be misleading and not reveal that the metadata for these records are incomplete.

- Introduced the "unspecifiedPart" element which may be used as an alternative to media-type specific parts (text, audio, video, etc.) when the media type value of a resource is not encoded in the original metadata[36].
- Added elements with free text values as an alternative to elements with controlled value vocabularies or combined elements that cannot be distinguished from the source metadata record (e.g., when size is encoded as a free text combining amount and size unit together)[37].

Finally, an important set of issues related to the documentation of licensing terms for resources has been resolved with the introduction of the "access rights" element and its usage together with the "conditions of use" element in the latest ELG release. More specifically, the following problems were observed for resources described by ELE informants but are also very common in catalogues:

- Total absence of a value for licence.
- Reference to a licence with multiple versions, without indicating the specific version.
- Reference to a licence by name and no further information on the licensing terms or a hyperlink to the licence text.
- Use of a free text value, such as "free for academic use", "available for research", etc.

The element "access rights" is assigned to the resource distribution as an optional element, or, in the relaxed version, as an alternative to the "licence" element and can be filled in with:

- a value from standard controlled vocabularies, such as the COAR access rights vocabulary[38], the rights statements of RightsStatements.org[39] and the ELE recommended values.
- a free text statement which is used for homogenizing standard licences without a version (e.g., "Creative Commons Attribution"), and for any value added by providers.

The "conditions of use" element is used as a filter in the faceted search of the catalogue. The values on this facet were selected in the previous release from a set of popular conditions of use associated with licences (e.g., attribution, non-commercial use, etc.) and were assigned, in the case of a subset of standard licences, by the ELG legal team, or, in the case of non-standard licences, by the metadata creators when they describe a resource. The assignment of conditions of use during this period has been extended to cover standard licences commonly used in the LRT sector and present in the ELG catalogue (see Section 4.2.2). The facet "conditions of use" has, therefore, been updated with a subset of these values, deemed important for findability purposes. In addition, the "access rights" currently populating the ELG catalogue have also been mapped to the same values, facilitating user search.

These changes have been deemed useful not only for the sources of this release but also for future imports from other catalogues that are populated with metadata records of interest to a broader range of communities

[36] It should be noted that wherever possible, the use of one of the ELG controlled values is the preferred option. Thus, for instance, the use of media type-specific values in other elements (e.g., format values such as DOC or WAV, which are used only for text or audio files, respectively) has been used to infer the media type in the case of crowdsourced metadata (cf. Section 5.2). In addition, in cases where the metadata records are imported from sources such as Hugging Face, where the vast majority of datasets are specific to a media type, we have opted for a default value.
[37] See https://european-language-grid.readthedocs.io/en/stable/all/A7_ReleaseNotes/releaseNotes.html#detailed-list-of-changes for a detailed list of changes.
[38] https://vocabularies.coar-repositories.org/documentation/access_rights/
[39] https://rightsstatements.org/en/

(e.g., Zenodo, EOSC, etc.) and the increasing use of more general metadata schemas, such as DataCite, DCAT and schema.org[40].

4.2 Legal Issues

4.2.1 Improvement of licence list in ELG

When we extended the export of Zenodo datasets as explained in Section 3.4, we realised that a number of licences were not part of the ELG metadata values. Thus, as part of our continuous work on Task 5.4, we asked the ELG legal expert 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 under other labelling, they were not used, or they had no link.

4.2.2 Addition of Conditions of Use in the ELG metadata

To improve the search functionality for resources based on their licensing conditions, it was decided to add a new metadata field corresponding to the "conditions of use" associated to each identified licence. For "standard" licences, the conditions of use were added automatically, based on information gathered from Creative Commons licences, values from the CLARIN licensing framework[41], META-SHARE licences, and the ELRA licence wizard[42]. However, for all other LRs, a thorough analysis was done by our legal expert.

During this process, a table with all SPDX licences was provided to the legal expert. This table displays 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.

During a first step of the process, the licences that were already in use within the ELG metadata scheme were thoroughly checked in relation with the conditions of use available. Afterwards, all licences provided in the SPDX table were checked in relation with these conditions of use. From this analysis three main categories of terms can be identified where to classify each licence:

1. The first category of conditions is linked with the rights granted by the licence, and it can be divided into the following rights. These conditions were essential to identify as they are the basic rights needed by ELG users to use the data made available on the platform.
    - Rights of reuse the data
    - Right to copy the data
    - Right to redistribute the data
    - Right to create derivatives
    - Right to sublicense the data
    - Grant of a patent licence on the patented application of data
2. The second main category of conditions that was identified is the one regarding the requirement imposed on the user on the redistribution and publication of the resources. These requirements are also

[40] https://schema.org/
[41] See https://www.clarin.eu/content/licenses-and-clarin-categories#res, https://www.clarin.eu/content/clarin-license-category-calculator
[42] http://wizard.elra.info/principal.php

necessary to users when dealing with the resource. These requirements were classified among the following conditions:

- Attribution requirement
- Obligation to document modification brought to the data or code
- Retention of the copyright notice attached to the original dataset
- Obligation to share the data under the same licence as the original dataset (ShareAlike)
- Obligation to share the data under an open licence or a licence compatible with the original licence (Copyleft requirement)

3. The third main category of conditions that was identified relates to the conditions imposed on users regarding the reuse of the data. Indeed, it was essential to know whether the data could be reused in certain applications. The conditions were split into the following items:

    - Commercial reuse of the data allowed
    - Commercial reuse forbidden
    - Evaluation use of the data
    - Academic use of the data
    - Language engineering research use
    - Research use
    - Training use

Other categories of conditions also include the requirement to identify both the user before allowing access and the type of user, however, during the analysis of the licences we discovered that these conditions were only mentioned in a few licences.

5 Gap Analysis for Language Resources

The European Language Grid is establishing itself as the entry point for thousands of datasets, tools, and services across Europe. In order to do so, ELG works in close collaboration with ELE, which has done an impressive job in identifying LRTs (see Section 5.2). This joint effort is crucial in their mission to identify and fill up the existing gaps. D5.2 already reported some gaps on certain resource types and certainly languages that have considerably improved with the insertion of the latest repositories as well as the ELE identification work.

Filling up gaps and contributing to improving current needs is also the objective of the Pilot Projects selected under the two ELG open calls for projects. Details can be seen further down in Section 5.1.

The following two sections describe the contributions mentioned. Section 5.3 will do a broad-coverage analysis of the current situation of the ELG catalogue in terms of datasets and provide some concluding statistics.

5.1 Contributions from the Pilot Projects

The ELG pilot projects are an excellent proof of concept for the ELG platform. D5.2 already introduced the Call 1 projects that could have a higher impact on both the leveraging and the enriching of the ELG catalogue. These projects have finished with the successful accomplishments that had been planned. As a result, a set of 52 da-

tasets have been inserted into the ELG catalogue, which are summarised in Table 3. This table also lists 2 already available datasets from two Call-2 projects. Furthermore, a large list of tools/services has also been contributed, which will be reported on in D4.3.[43]

Project	Dataset
European Clinical Case Corpus (E3C)	European Clinical Case Corpus (2.0.0)
	European Clinical Case Corpus - raw version (1.1.0)
Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools for the European Language Grid (EVALITA4ELG)	SardiStance Dataset (1.0.0 (automatically assigned))
	ATE_ABSITA dataset (1.0.0 (automatically assigned))
	SardiStance Test Set (1.0.0 (automatically assigned))
	SardiStance Training Set (1.0.0 (automatically assigned))
	AMI 2020 Dataset (1.0.0 (automatically assigned))
	SardiStance Gold Labels (1.0.0 (automatically assigned))
	AMI 2018 Dataset (1.0.0 (automatically assigned))
	DADOEVAL corpus (1.0.0 (automatically assigned))
	Cross-Genre Gender Prediction (GxG) dataset (1.0.0 (automatically assigned))
	EVALITA 2007 Parsing Task Dataset (1.0.0 (automatically assigned))
	CONcreTEXT - Concreteness in Context (1.0.0)
	DIACR-Ita dataset (1.0.0)
	IronITA (1.0.0)
	CHANGE-IT dataset (1.0.0)
	EVENTI corpus (1.0.0)
	ABSITA dataset (1.0.0)
	ITAmoji dataset (1.0.0)
	iLISTEN dataset (1.0.0)
	SENTIPOLC 2016 dataset (1.0.0)
	SENTIPOLC 2014 dataset (1.0.0)
	PoSTWITA dataset (1.0.0)
	HaSpeeDe 2018 dataset (1.0.0)
	HaSpeeDe 2 Dataset (1.0.0)
	QA4FAQ Dataset (1.0.0)
	Dependency Parsing Dataset (1.0.0)
	CRIPCO corpus (1.0.0)
	FactA Dataset (1.0.0)
	EVALITA 2011 Parsing Task Dataset (1.0.0)
	ArtiPhon Dataset (1.0.0)
	Anaphora Resolution Dataset (1.0.0)

[43] The ELG Deliverable D4.3 “Services, Tools and Components (final release)” is due at M37 (January 2022).

	PRELEARN Dataset (1.0.0)
	DankMemes Task A Dataset (1.0.0)
	DankMemes Task B Dataset (1.0.0)
	DankMemes Task C Dataset (1.0.0)
	DankMemes Dataset (1.0.0)
	TAG-it Dataset (1.0.0)
	AcCompl-it Dataset (1.0.0)
	Dependency Parsing Task Dataset (1.0.0)
	Constituency Parsing Task Dataset (1.0.0)
	Textual Entailment Dataset (1.0.0)
	Lexical Substitution Task Test Set (1.0.0)
	SUGAR Dataset (1.0.0)
	KIPoS corpus (1)
Open Translation Models, Tools and Services (OpusMT)	OPUS-MT: Celtic-English machine translation model (1.0.0 (automatically assigned))
	OPUS-MT: English-Celtic languages machine translation model (1.0.0 (automatically assigned))
	OPUS-MT: English-Indo-European languages machine translation model (1.0.0 (automatically assigned))
	OPUS-MT: English-Multiple languages translation model (1.0.0 (automatically assigned))
	OPUS-MT: Indo-European languages-English machine translation model (1.0.0 (automatically assigned))
	OPUS-MT: Indo-European languages-Indo-European languages machine translation model (1.0.0 (automatically assigned))
	OPUS-MT: Multiple languages-English machine translation model (1.0.0 (automatically assigned))
Turku Paraphrase Corpus (TurkuParaC)	Turku Paraphrase Corpus (1.0.0 (automatically assigned))
CEFR Labelling and Assessment Services (CEFRSERV)	Multilingual CEFR Word List (1.0.0 (automatically assigned))

Table 3: Datasets contributed by the ELG Pilot Projects

5.2 Collaboration with ELE: Receiving input from Crowdsourcing Initiatives

ELG collaborates with the European Language Equality (ELE) project[44] in order to promote digital language equality in Europe. ELE has launched a survey[45] to collect information on language resources and technologies available for the languages under investigation. For this purpose, a web form has been designed soliciting information for records with the following metadata elements:

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

- For all LRTs: Resource name, Resource short name, Landing page, Resource description, Resource type, (Funding) project(s), Funding type, Keywords, Resource publication year, Licence or Type of access, Resource provider (organisation), Source of information, Contact email (with Surname and Name (for persons)).
- In addition, for data resources: Subclass, Language(s), Language geographical variety, Linguality Type, Multilinguality type, Media type(s) of parts, Size, Domain(s) and:
  - for corpora: Annotation type(s), and,
  - for lexical/conceptual resources: Encoding level(s).
- Additionally, for tools/services: Function(s)/Task(s), whether they are Language independent, and if not, Language(s) of input and output, Language geographical variety, Language(s) of output, Media type(s) of input and output, and TRL.

The web form was made available to the ELE consortium partners that are responsible for the collection of information on LRTs and reporting on the status of the language ("informants"). The results of the web form are exported in a tabular format. The aggregation of the results (about 6,500 records) has been the input for import into ELG. However, before the conversion and final import, a long process of normalisation and curation was necessitated. This was due to several factors such as:

- Elements with values taken from a controlled vocabulary: restrictions are hard to impose on web forms, especially for long lists of values or for elements that take multiple values. For instance, although a set of language values was offered for selection on the web form, informants could also add other values, which resulted in values with typos, unofficial names, alternative names, etc., which had to be normalised and mapped to the ISO 639 language codes.
- Mandatory elements that have not been filled in: Elements that are mandatory in the ELG schema but have not been filled in in the ELE forms[46] have been addressed in different ways, depending on various factors, such as importance, dealing with similar cases in the future, etc.:
  - As ELG is already in the process of importing metadata records from general purpose catalogues[47] a relaxed version of the ELG schema was introduced in this release (see Section 4.1). Thus, in order to address elements such as "media type", which are considered important for ELG but are not always present in metadata records of general catalogues, we have introduced a value "unspecified" which is allowed only for metadata records imported into ELG from bulk imports.
  - For elements such as "licence", which is of utmost importance for the use of a resource in a clear legal context[48], but which is difficult to find, especially for legacy resources, we have again relaxed the schema and introduced the "access rights" element with recommended values to indicate level of restriction (for all or some uses, with a fee)[49] .

[46] The collection of information on resources from ELE informants was a time and effort-consuming endeavour, with many diverse sources of information (specialised catalogues, general catalogues, institutional websites, etc.), which include different types of information and of a different level of detail. Due to time restrictions, a thorough investigation of the metadata elements of all resources would not have been realistic. Therefore, the selection of elements added in the ELE form, and its design aimed to facilitate this endeavour and support both the manual completion for single metadata records or the batch addition of metadata records. The reasons behind this choice were the identification of resources for all target languages to the extent possible and their aggregation in a single source (ELG catalogue).
[47] Cf. Sections 3.4 and 3.6 on Zenodo and Hugging Face, respectively.
[48] See also FAIR principles: https://www.go-fair.org/fair-principles/r1-1-metadata-released-clear-accessible-data-usage-license/
[49] It should be noted though that even in such cases, more values were introduced by the informants and had to be normalised before the import into the ELG platform.

- Limitations of the tabular format: Although the CSV format presents some advantages, given its simplicity and users' familiarity, and the fact that it was used behind a user-friendly web form, it still poses many difficulties for validation purposes, especially for elements with patterns, or with multiple values. For instance, the email element was filled in with free text values, URL links, etc., since no validation pattern was used for it. For elements with multiple values, such as languages, functions, etc. different delimiters were used and had to be normalised. The two cases that posed the most problems in this category were those of language and project, because the information for them was split in two columns: language and region, for the former, and project and funding type for the latter. For these cases, we had to check and ensure that the same number of values was consistently used across the two complementary columns and, moreover, that the values were correctly matched.
- Presence of duplicate metadata records: Exact and near duplicate metadata records have been identified during the normalisation process in the survey results. These duplicates are due to various reasons: the addition of the same resource by different informants or from different sources, the use of the same name or short name for related resources (e.g., a corpus and its annotated version, a lexicon, and its access software), different versions of the same resource, etc. At this phase, we have only looked into cases of resources with the same name, short name or identifier (which are also the elements used for identifying duplicates in the ELG platform). However, the identification of exact and near duplicates will be undertaken in a curation process to be determined at a later stage and to be performed for all metadata records in the ELG platform.

As a result, **4,127 metadata records for data resources and 2,215 for tools** have been imported into ELG. The metadata records can be claimed[50] by the resource creators and enriched with further information. For this reason, e-mail messages have been sent to all contact persons included in these metadata, notifying them of their publication in the ELG catalogue. The process of claim by the resource owners and assignment to them is ongoing. Further dissemination steps will be taken up at later stages to encourage this procedure.

5.3 Analysis of the ELG Content and Detected Gaps

Following the latest ingestions of datasets as well as the contributions from the Pilot Projects and the ELE identification, the ELG catalogue has reached a total of 8873 metadata entries, which is a very good achievement so far[51]. Most of these entries are description records without the data being hosted in ELG (103 resources are fully available through the platform). However, most of the ingested datasets are available in the referenced repository page, often available for download, and this is reflected in the ELG catalogue too.

Regarding resource types and their linguality, the numbers are summarised in Table 4 and illustrated in Figure 3. As expected, the highest numbers go for the corpora (6,236 of them in the ELG catalogue), with twice as many monolingual corpora in comparison with bilingual ones (which in turn are three times as many as the multilingual ones). Lexical/Conceptual resources are also very well represented with 2229 entries.

One of our bigger concerns when we wrote the D5.2 report was the fact that there were barely any language descriptions (there were only 7). This has changed with the work towards R3 and by now, we count 408 of them, the majority of them being monolingual.

[50] See https://european-language-grid.readthedocs.io/en/stable/all/3_Contributing/Claim.html for the claim procedure.
[51] Details on the breakdown of sources for these 8873 can be seen in Figure 1.

Further on the language descriptions, the new number for the "language models" subclass (cf. Table 5) has increased to 358. This is very good news as models are very popular and highly demanded resource types. It is a positive outcome, and they provide a good addition and further development for the ELG catalogue. ELG has encouraged the use of its platform for the creation of models and one particular Pilot project (OpusMT) has supported this resource type by contributing seven models (see Table 3).

Resource Types and their Linguality	Number of Resource Types
Corpus	6236
bilingual	1850
monolingual	3778
multilingual	604
N/A	4
Language description	408
bilingual	26
monolingual	339
multilingual	43
Lexical/Conceptual resource	2229
bilingual	680
monolingual	1138
multilingual	411
Total	8873

Table 4: ELG Content – Types of resources split according to linguality

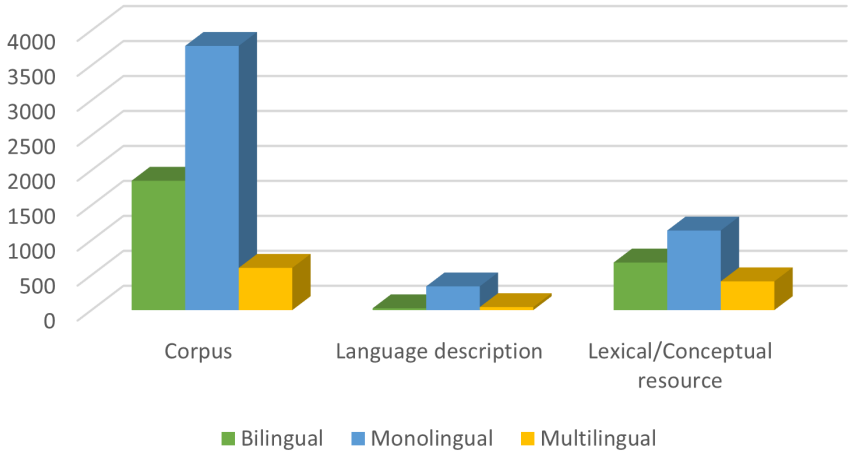

Figure 3: ELG Content – Types of resources split according to linguality

Resource Types and Subclasses	Number of Resource Types
Corpus	6236
raw	2873
annotated	2477
annotations corpus	4
unspecified	882
Language description	408

models	358
grammars	46
others (descriptions)	4

Table 5: Resource subclasses for corpus and language description

Besides illustrating the ingestion status of language models, Table 5 also shows the distribution of the corpus resource type among different subclasses (mostly *raw* and *annotated*), with very similar figures for both.

6 Experience in using ELG datasets for Model Training

This section is a return on experience leveraging the ELG datasets for model training under T5.3 (task reported in D5.2 and concluded in M24). As mentioned in the previous section, ELG encourages and supports the use of its platform for the creation of highly demanded datasets such as language models. For this task to be a useful and user-friendly activity, it is important to know what can be done, how, under what conditions and limitations and whether the output results reflect the user's expectations. This sort of user validation is very relevant for future improvements.

As discussed in D5.2, we focused in T5.3 on using the ELG to identify suitable datasets for training models for neural machine translation. To this end, we developed python software to access the ELG interface programmatically (also reported in D5.2). We also contrasted the experience of a potential user of the ELG with existing alternatives, most notably the OPUS collection of corpora[52] and the software package mtdata[53].

We can summarise our experience and progress as follows:

- All three approaches to providing access to training data for machine translation are currently actively maintained. OPUS and mtdata focus specifically on parallel data, ELG, on the other hand, aims for much broader coverage of language resources. This means that the ELG search interface should make specific provisions for searching for parallel data with specific filters for usage conditions. This is especially the case to discern whether commercial use is allowed. Of the three packages mentioned, to the best of our knowledge ELG is the only one that provides a search interface that allows to filter for specific conditions of use. On the other hand, not all properties of registered resources are returned by default by the ELG search interface; our software currently mitigates this by requesting full metadata records for each record retrieved through the search interface (in order to update the cache), but in the long run this is inefficient.
- There is a risk of data duplication loops. For example, OPUS resources are currently automatically ingested by ELG, whereas OPUS also appears to be monitoring web sites such as elrc-share.eu (which in turn is also ingested periodically by ELG). To make things worse, there are also other projects such as nteu.eu that actively collect publicly available training data and publish their training data in repositories monitored by the ELG in aggregated format, leading to potential data duplication that is not obvious at first sight.

[52] https://opus.nlpl.eu
[53] https://github.com/thammegowda/mtdata

- Most datasets that we investigated came with their own set of problems, be it data encoding errors, poor sentence alignments for allegedly parallel data, etc. This is a problem that does not affect the ELG exclusively. Quite the opposite, proper data curation may offer a future business opportunity for the ELG, as data curation is at the same time crucial for good systems, tedious, and expensive. Especially for EU/EC-sourced data (where usage rights may be more straightforward), active data curation, as said above, may constitute a business opportunity for the ELG. As the ELG grows (and is able to afford more compute capacity through incoming revenue), active quality estimation for uploaded datasets might be an attractive service to be offered by the ELG.

7 Summary and Conclusions

D5.3 has reported on the work done since D5.2 (R2) with regard to the provision of datasets to the ELG catalogue. The platform is at a very advanced stage by now and it hosts 8,873 datasets from several repositories and contributions. ELG goes beyond the concept of the basic catalogue by handling analysis, conversion, mapping, harvesting and description aspects, depending on the source to ingest. ELG has advanced very positively since R2, a) with the ingestion of many more datasets (we have gone from 2737 to 8873), b) with the establishment of perennial procedures such as harvesting protocols to continue feeding the catalogue with updates from already processed sources, c) with the definition of clear legal frameworks and licensing schemas to protect our sharing activities, d) with the setting up of collaborations (such as ELE) that move forward for a common target, delivering language resources and services to the community. ELE's identification work has boosted the findings and integration for ELG and has contributed considerably towards the filling up of gaps, a still future challenge, in particular for under-resourced languages.

Future work will continue looking into enriching the catalogue for the users and doing it in an intelligent manner as we learn from users' needs and experience. The feedback described on the model training experience will be taken into account for the future months or beyond.

A. Annex 7. Mapping of elements Hugging Face – ELG

HF element	ELG element	Datatype (ELG)	Comment
pretty_name	resource name	free text	if it exists, preferred over id
id	resource name	free text	
description	description	free text	
thumbnail	logo	link	
url of the HF view page	landing page	URL	
task category	intended application	controlled vocabulary	values mapped
task id	intended application	controlled vocabulary	values mapped
benchmark	keyword	free text	
task category	keyword	free text	if the value has not mapped to intended application

task id	intended application	controlled vocabulary	if the value has not mapped to intended application
type	keyword	controlled vocabulary	
citation	is to be cited by	complex element	a parser for bibtex records is used to map the record into the distinct elements in the ELG element (e.g., author, title, publisher, etc.)
annotations_creators	corpus subclass	controlled vocabulary	if the annotators_creators element exists, the value is “annotated”; otherwise, default value is “raw”
–	mediaType	controlled vocabulary	default value is “text”
language	language id	controlled vocabulary	values mapped
–	dataset distribution form	controlled vocabulary	default value is “downloadable”
url of the HF view page	access location	URL	
size	–	complex element	cannot be mapped because HF is a free text
–	data format	controlled vocabulary	default value “unspecified”
licence	licenceTermsName	free text	values mapped from SPDX; if not in SPDX list, searched for the relevant URLs and added separately; if the URL is not found, the resource is not imported