Chapter 2
The European Language Grid Platform: Basic Concepts

Stelios Piperidis, Penny Labropoulou, Dimitris Galanis, Miltos Deligiannis, and Georg Rehm

Abstract In the fragmented Language Technology (LT) landscape of multilingual Europe, ELG has set out to bring together language resources and technologies (LRTs) and boost the LT sector and its activities. The primary goal is to build a scalable and comprehensive cloud platform for providers, developers, integrators and consumers of language resources and technologies. We describe the basic concepts of the ELG platform in terms of its architecture, the functionalities and services offered to its types of users and the policies it implements. We present the ELG repository, its catalogue features, the LT services execution environment as well as the metadata model underlying the platform operations and the resources life cycle, from creation to publication. We also discuss the compliance of ELG with the FAIR principles and the relation to other platforms and infrastructure initiatives which have inspired certain aspects and with which ELG has been establishing strong links.

1 Introduction

The overarching objective of the European Language Grid (ELG, Rehm et al. 2021) is to tackle the observed fragmentation in the European Language Technology (LT) landscape by bringing together Language Resources and Technologies (LRTs), commercial and non-commercial, and through multiple multi-level services support and boost the LT sector and LT activities in Europe. The primary technological goal is to build a scalable cloud-based platform through which developers and providers of language resources and technologies can not only deposit and upload their resources and technologies into ELG, but also deploy them through the platform and make use of the services, technologies and resources made available by others. ELG is a marketplace through which consumers and integrators of LRTs can discover, try out

Stelios Piperidis · Penny Labropoulou · Dimitris Galanis · Miltos Deligiannis
Institute for Language and Speech Processing, R. C. "Athena", Greece, spip@athenarc.gr, penny@athenarc.gr, galanisd@athenarc.gr, mdel@athenarc.gr

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_2

and integrate the resources and technologies they require for their own research and application development.

The primary services of the platform are dedicated to the deposition, discovery, distribution and deployment of Language Resources and Technologies. ELG already offers access to thousands of commercial and non-commercial LTs and ancillary LRs for all European languages and more. These include processing and generation services, tools, applications for written and spoken language, as well as datasets, corpora, lexical resources, language models and computational grammars.

ELG also supports the promotion and collaboration of LT stakeholders through an extensive catalogue of organisations (companies, SMEs, academic and research organisations and groups, etc.) active in the LT community. Organisations can describe, promote and distribute their services and resources all in one place. Complemented with an expanding catalogue of European and national projects that have funded the production of LRTs and related activities, the catalogue of the ELG platform offers an overview of the European LT landscape. ELG, therefore, also acts as an observatory of LT, consolidating existing and legacy tools, services, LRs, and information about them, as well as newly emerging ones. This, in turn, enables the identification of gaps and imbalances between the LRTs offered for all European languages, a valuable instrument for the support of digital language equality in Europe.

ELG is conceived as a platform for the whole LT community. Primarily for Europe, ELG is a platform built *by* the European LT community *for* the European LT community, including industry, innovation and research. For the population of the catalogue of its platform, ELG builds bridges to existing initiatives and reaches agreements for harvesting and importing information (i. e., metadata) and resources from other infrastructures, platforms and repositories under mutually agreed conditions, business policies, acknowledgement and attribution of the source, and collaborates in joint initiatives and crowdsourcing campaigns.

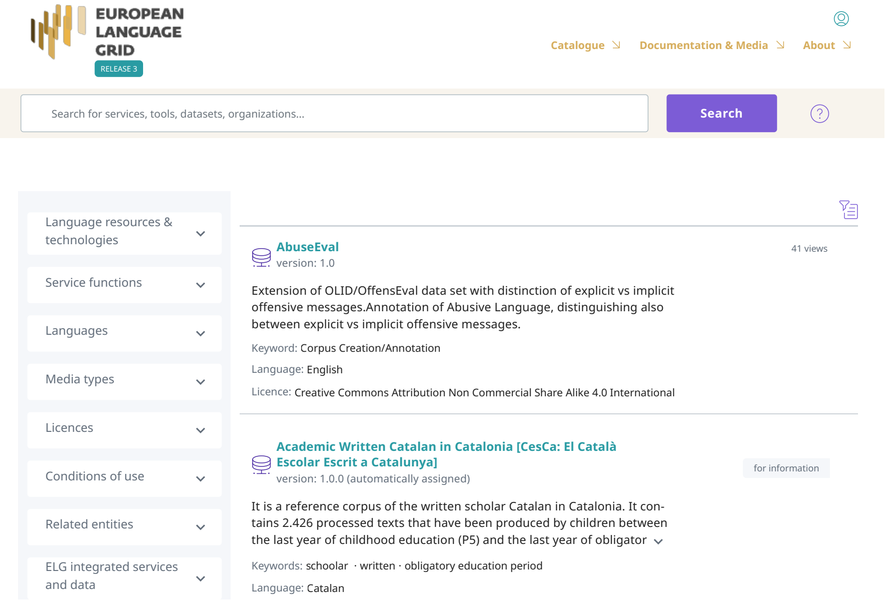
This chapter introduces the basic concepts of the ELG platform, while the subsequent chapters go into more detail with regard to functionalities offered to consumers (Chapter 3) and providers (Chapter 4), the cloud infrastructure (Chapter 5) and the synergies with other initiatives (Chapter 6). We first give an overview of the platform features (Section 2) and its users (Section 3). Section 4 presents the architecture of ELG. Sections 5 and 6 present the models and policies that influence the design and operations of the ELG platform, i. e., the metadata model, and the publication life cycle of catalogue entries. Section 7 positions the ELG platform with regard to the FAIR principles (Wilkinson et al. 2016).

2 Overview of the ELG Platform

The ELG platform combines the features of a catalogue (Section 2.1), a repository (Section 2.2), and an execution environment for running services (Section 2.3).

2.1 Catalogue

All LRTs are accessed through their metadata records in the catalogue (Figure 1). Providers can describe and share their LRTs; they can upload them to be hosted in ELG, or they can only describe them and provide access to them through other locations, such as institutional or national repositories, or private repositories of commercial organisations. They can also create dedicated pages for their organisations, describe their offerings and services and interlink all their LRTs through their own pages.



AbuseEval
version: 1.0

Extension of OLID/OffensEval data set with distinction of explicit vs implicit offensive messages.Annotation of Abusive Language, distinguishing also between explicit vs implicit offensive messages.

Keyword: Corpus Creation/Annotation
Language: English
Licence: Creative Commons Attribution Non Commercial Share Alike 4.0 International

Academic Written Catalan in Catalonia [CesCa: El Català Escolar Escrit a Catalunya]
version: 1.0.0 (automatically assigned)

It is a reference corpus of the written scholar Catalan in Catalonia. It contains 2.426 processed texts that have been produced by children between the last year of childhood education (P5) and the last year of obligator

Keywords: schoolar · written · obligatory education period
Language: Catalan

Fig. 1 Browse/Search page of the ELG catalogue

Additionally, the ELG catalogue includes metadata records imported automatically from other sources, through standard harvesting protocols and dedicated converters, thus resulting in an extensive and continuously growing inventory of LRTs as well as of organisations and projects in the LT domain.

LRT consumers, i. e., users, and other interested parties can search for and discover LRTs using free text search and faceted views of the catalogue. Users can select and view the detailed descriptions of LRTs to see if they fit the users' needs. Users can access the resources, either directly if hosted in ELG, or be re-directed to the URL from where the resources are accessible. Users can also search for organisations, browse them, and view their activities on their profile pages. If these organisations have also described the LRTs they developed, users can navigate to the respective pages for more details. Last, users can also discover the LT-related projects in which organisations participated and that have helped fund the organisations' LRT development. Finally, users can export and download the metadata descriptions or share the pages on social media.

2.2 Repository of Language Resources and Technologies

LRT providers can upload their resources to be hosted in the ELG cloud infrastructure, and to be made available to consumers for direct download. Providers must specify the licensing conditions under which the resources can be used. Depending on the terms, ELG will allow immediate download (for open access resources) or impose further measures (authentication and authorisation). Commercial LRTs, distributed for download at a fee, will be available for purchase using a user-friendly billing service.

ELG as a repository is committed to making data, services and their metadata FAIR, i. e., findable, accessible, interoperable and reusable (Wilkinson et al. 2016). The assignment of persistent identifiers in the form of Digital Object Identifiers (DOIs)[1] for the data and services hosted in ELG is among the main steps towards this objective; the FAIR principles, detailed in Section 7, form an integral part of the ELG policies aiming to support the requirements posed by research results reproducibility objectives and practices.

2.3 Running Language Technology Cloud Services

To benefit from the advanced features of ELG, providers can integrate LT tools as ready-to-deploy services, following our specifications (Chapter 4). In this case, consumers can test the tools and services using the trial UIs or APIs offered by ELG, and, ultimately, integrate them in their workflows and systems. For commercial services, billing services will be available to allow pay-for-use services with seamless access and use in the minimum possible number of steps.

ELG provides a set of standard APIs which cover all principal service types (see Chapter 3, Section 3, p. 50 ff., for more details): *information extraction and annotation services* for *text and speech*, *text-to-text* services (most notably machine translation services, but also summarisers, anonymisers, etc.), *classification services for text or image*, such as language identifiers, fake news detectors, sentiment analysers, etc., *speech recognition* services, *text-to-speech synthesis* services, and *image OCR* (optical character recognition) services.

The technical specifications give service providers a set of easy-to-implement integration options from which they can select the one that best fits their needs. All that is required is that they upload an image of their tool or service using one of these options in a container registry and provide access to ELG.

[1] https://www.doi.org

ELG maintains a dedicated container registry for LT services.[2] As the images of LT services are partly pulled from registries external to the ELG project, this registry serves as a point to collect LT service images when they are ingested into the ELG and to apply versioning. This approach enables us to ensure that older versions of images remain available even if their original site no longer provides them.

To provide easy access and interaction with the ELG platform also for programmers, a Python SDK has been developed on top of the various ELG programmatic interfaces providing simple methods to easily interact with the platform and consume resources in Python (see Chapter 3, Section 4, p. 55 ff., for more details).

3 User Types and User Model

Specified by its mission, ELG targets various types of users, broadly classified into:

Providers of LRTs, both commercial and academic, albeit with different requirements (the former seek to promote and sell their products and activities, while the latter wish to make their resources available for research or look for cooperation to further develop them in new projects or commercialize them),

Consumers of LRTs, including companies developing LT tools, services and applications, integrators, researchers using LRT for their studies, etc.,

LT laypersons interested in finding out more about LT and its uses,

Funding authorities and stakeholders that wish to get an overview of the LT field and landscape, trends and prospects with regard to languages, domains etc.

All users can browse the catalogue and access, view and inspect the detailed descriptions of the assets listed in the catalogue, and download resources available with open access licences. For further interactions with the ELG platform, registration is required and can be performed with a simple and user-friendly self-service procedure. The types of permitted actions and access level are determined by the user role: *registered consumers* can run integrated services and download resources that are available for free download to authenticated users; *providers* can, in addition, describe all types of assets, upload content files, and integrate services according to the ELG technical requirements; two specific user roles (*validator* and *administrator*) are reserved for ELG team members responsible for the management of the catalogue, metadata records and data files, in accordance with the ELG policies (Section 6) including the overall platform maintenance and administrative operations.

[2] registry.european-language-grid.eu

4 Architecture

The ELG platform uses state-of-the art technologies and is designed to evolve over time to address new requirements or technological advancements. The choices made in the architectural design and implementation allow for scaling with the growing demand and supply for compute resources and lay the foundation for interoperable data and service spaces.

All subsystems are built with robust, scalable, reliable, widely used open source technologies, as described below. Docker containers[3] are used for all services and applications which comprise the ELG platform, while Kubernetes[4] is used for container orchestration. Conceptually, ELG takes the form of a three-layered platform, with each layer grouping together the main subsystems responsible for the platform's functionalities: *base infrastructure*, *platform back end*, *platform front end* (Figure 2).

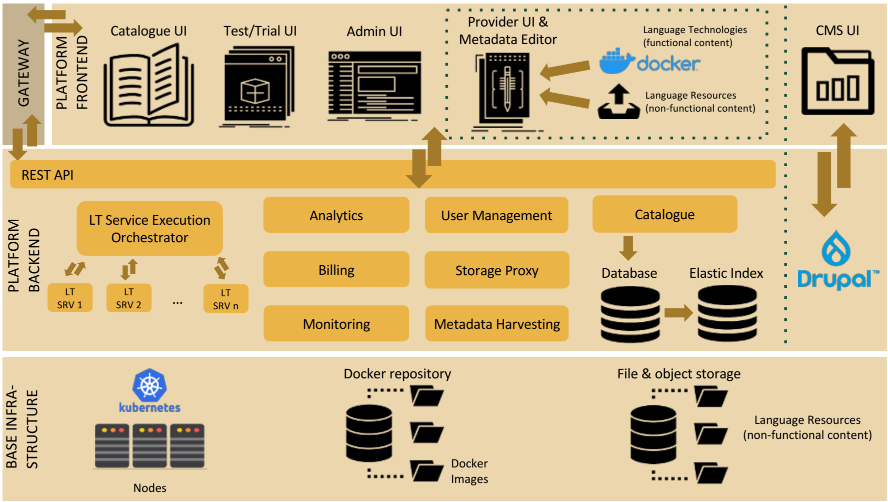


Fig. 2 ELG platform architecture

The *base infrastructure* is the layer on which all ELG software components are deployed and run. It includes the supporting tools that facilitate development and management of the ELG platform software. It is composed, first and foremost, of the compute nodes running the platform, alongside their respective volume storage and networking facilities; these are organised in two different clusters, one for development and one for production purposes. It also comprises public and private

[3] A Docker image of an application contains its actual code and all required dependencies required to run it; e. g., the operating system, frameworks, settings, configuration files, libraries, etc. Containers are instantiations of images and can be thought of as lightweight virtual machines.

[4] Kubernetes is a framework that enables and simplifies the deployment, scaling and management of containers, see https://kubernetes.io.

container registries, which host all images for the ELG platform components and for the LT services integrated in the platform. In addition, it includes an S3-compatible file and object storage, through which data resources uploaded by providers as well as backups of core platform components are persisted. This layer also includes a set of Git[5] repositories for the source code of the platform software apps and for the individual LT services implementations of specific providers. Chapter 5 (p. 95 ff.) provides more information on the base infrastructure.

The *platform back end* consists of all the components that enable the operation of the ELG platform, i. e., the catalogue core components, the component for processing LT services and platform support as well as management components. The catalogue component, implemented using Django[6], interfaces with a PostgreSQL[7] database for storing the metadata records and an index, which uses ElasticSearch[8]. The LT service execution server offers a common REST API for calling LT services integrated in the platform, and handles failures, time-outs, etc. Finally, separate modules are used for the user management and authentication module (based on Keycloak[9], an identity and access management solution), the analytics, monitoring, metadata harvesting and the proxy for interacting with the S3-compatible storage.

The *platform front end* layer consists of the static pages maintained in a Content Management System (CMS). These provide information on the ELG project and initiative, and the platform UIs for the different types of users, i. e., consumers, providers, validators, and administrators. These include the catalogue pages (browse, search, view), and the dashboard pages customised for the different user types, UIs for registering (describing and uploading) LRTs and other assets and supporting the publication life cycle, implemented using React[10], and the trial UIs for services integrated in ELG. The catalogue UI consumes REST services exposed by the ELG platform back end (e. g., catalogue application, LT Service execution server).

Chapters 3 (p. 37 ff.) and 4 (p. 67 ff.) provide more information on the back end and front end layers of the European Language Grid platform.

5 Catalogue Contents and Metadata Model

All types of LT assets as well as all LT-related meta-information are brought together, aligned and interlinked. This set of information[11] is formally structured and harmonised in ELG using the ELG-SHARE metadata model[12] catering for the full

[5] https://git-scm.com

[6] https://www.djangoproject.com

[7] https://www.postgresql.org

[8] https://www.elastic.co

[9] https://www.keycloak.org

[10] https://reactjs.org

[11] https://european-language-grid.readthedocs.io/en/stable/all/A2_Metadata/Metadata.html

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

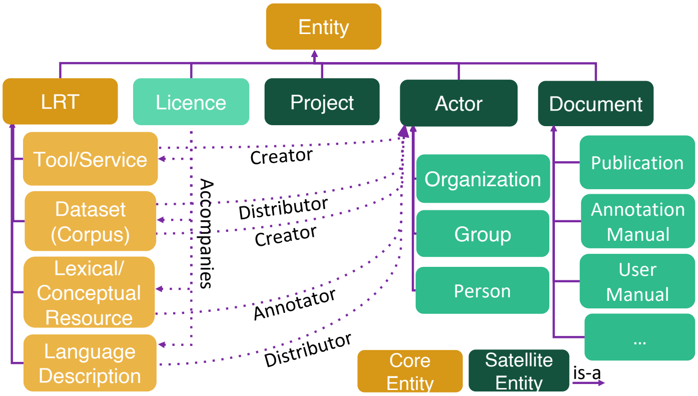


Fig. 3 ELG entities

language data and services life cycle and their related entities (Labropoulou et al. 2020). The ELG model covers the following types of entities (Figure 3).

- *Language resources and technologies (LRTs)*, further classified into:
  - *Corpora*, i. e., datasets of mono/bi/multilingual text documents, audio/video recordings, multimedia datasets, parallel corpora, translation memories, etc.
  - *Lexical/conceptual resources*, including lexica, ontologies, gazetteers, term lists, computational dictionaries, etc.
  - *Language descriptions*, which mainly refer to computational grammars, statistical and machine learning models
  - *Tools/services*, i. e., pieces of software offered as locally executable code or web services, hosted and running in the ELG cloud platform or remotely
- Related/satellite entities, such as actors, be it *persons* or *organizations* that have created or that curate resources, *projects* that have funded them or in which they have been used, as well as *licences* and accompanying *documents* (e. g., publications related to the resource, user manuals, technical documents, etc.)

The ELG model lies at the heart of the platform and supports its key operations. In particular, it aims to 1. support the discoverability of all catalogue contents; 2. enable accessibility by human users and, where possible or required, machines (e. g., including links to URLs that offer direct access to a resource or service); 3. address (at the metadata level) interoperability requirements of resources belonging to the same types and media, but coming from different sources with different descriptions, as well as between resources of different types and media (e. g., between datasets and services to be used for their processing); and, 4. finally, satisfy documentation needs at different levels of granularity, ranging from the strict enforcement of technical metadata required for the deployment of ELG-compatible services to rather loose descriptions of resources imported from general purpose catalogues.

The metadata model builds upon previous work from the META-SHARE metadata model (Gavrilidou et al. 2012), which caters for the description of language resources and language-processing technologies, and its application profiles, i. e., ELRC-SHARE (Piperidis et al. 2018a), OMTD-SHARE (Labropoulou et al. 2018), CLARIN-SHARE (Piperidis et al. 2018b), which extend, restrict and adapt the basic model to specific domains and areas (e. g., public domain resources, text and data mining domain, etc.), and the MS-OWL ontology[13] (McCrae et al. 2015; Khan et al. 2022), which is the RDF/OWL representation of the model.

The model builds along three key concepts, each of which is associated with a distinctive set of metadata elements:

- *resource type*, with the four subtypes described above;
- *media type*, which specifies the form or physical medium of the resource. The notion of medium is preferred over the written, spoken or multimodal distinction, as it has clearer semantics and allows us to view LRs as a set of modules, each of which can be described through a distinctive set of features. Thus, the following media type values are foreseen: *text*, *audio*, *image*, *video* and *numerical text* (referring to numerical data, such as biometrical, geospatial data, etc.). To cater for multimedia and multimodal language resources (e. g., a corpus of videos and subtitles, or a corpus of audio recordings and transcripts, a sign language corpus with videos and texts, etc.), language resources are represented as *consisting* of at least one media part;
- *distribution*, which, following the DCAT[14] model (Albertoni et al. 2020; Maali and Erickson 2014), refers to any physical form of the resource that can be distributed and deployed by end-users.

These elements give rise to a modular structure, in which metadata elements are attached to the appropriate level ("class"). The "LanguageResource" class includes properties common to all resource and media types, such as those used for identification purposes (title, description, etc.), recording provenance (creation, publication dates, creators, providers, etc.), contact points, etc. More technical features and classification elements differ across resource and media types and are, thus, attached to combinations thereof; for example, a corpus may take elements specific to annotation processes, while the description of a computational lexicon encodes, e. g., whether it includes lemmas, examples, grammatical information, translation equivalents, etc. Technical features, such as format, size, information on licensing and mode of access are properties of the distribution. They can also differ across resource type. For example, corpora can be distributed as PDF files or as simple text files, lexical resources in tabular form or queried through an interface, while tools may be available as source code, executable files or web services. Each of these forms can be licensed under different terms: source code may be available at a price for integration in other applications, while an API may be offered for research purposes without any fee. Figure 4 illustrates a subset of the elements for a tool/service.

[13] http://w3id.org/meta-share/meta-share

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

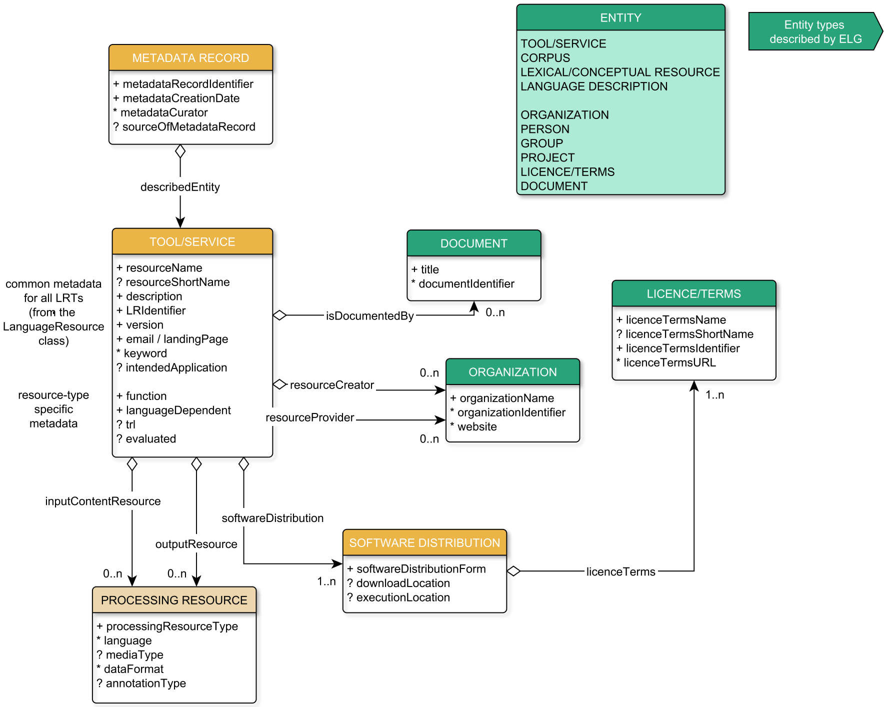


Fig. 4 Excerpt of the minimal schema for tools/services

The schema allows for the description of the full life cycle of language resources (see, e. g., Rehm 2016), from conception and creation to integration in applications and usage. All this information leads to a complex and demanding schema; to ensure flexibility and uptake by resource providers, the elements are classified into three levels of optionality:

- *mandatory:* elements that are necessary
- *recommended:* elements that can help the current or future use of the resource, or useful information that providers have not yet standardised
- *optional:* all remaining information

The minimal schema comprises all mandatory elements which must be filled for a metadata record to be considered ELG-compliant and eligible to be registered in the platform. Recently, a "relaxed" version of the ELG schema was introduced as a way of handling metadata records with "lighter" information imported from other catalogues in ELG, but this version of the schema is allowed only under specific circumstances. Chapter 6 discusses this in more detail. Below, we summarise the metadata categories considered mandatory for the description of resources (Figures 6 to 10 in the Appendix provide an overview for each resource type).

- Administrative information: these features are important for the identification of an LRT (resource name, version, description which includes information on the contents, provenance information, any other information deemed useful and helpful for consumers, etc.), contact information (landing page with additional information or a contact email).
- Classification information: one or more free text keywords that support the findability of the resource.
- Usage information: separate distributions for each distributable form of the resource, with the following elements: the distribution form (i. e., whether it can be downloaded, accessed through an interface, deployed as a web service, etc.), the licensing terms under which it can be used (licence name and URL); if the resource is not uploaded in ELG, an access or download link.
- Legal/ethical information for data resources: whether personal or sensitive data is included and, if applicable, information on anonymisation.
- Technical information: depending on the resource type
  - for tools/services: the function (i. e., the task it performs, e. g., named entity recognition, machine translation, speech recognition, etc.), the technical specifications of its input (at least the resource type it processes, e. g., corpus, text, etc.), whether it is language independent and, if not, the input languages; depending on the function, further information may be required (e. g., the languages of the output resource for machine translation services);
  - for all data resources[15]: features on the language following the BCP 47[16] guidelines, multilinguality type, resource subtype with different values (e. g., terminological glossary, ontology, etc. for lexical/conceptual resources, raw or annotated for corpora); size and format information must also be added separately for each distribution and media part;
  - in addition, specifically for models: the intended application (e. g., machine translation, named entity recognition, etc.), the model function (e. g., zero-shot classification), and model type (e. g., embeddings, Bayesian model, n-gram model, etc.);
  - specifically for grammars and lexical/conceptual resources: the encoding level of their contents (i. e., whether they contain morphological, syntactic, semantic, etc. information).

For organisations and projects, all that is required is the name (official title). However, we also recommend a free text description with the activities of the organisation or the project summary respectively, and the URL of its website. The LT area(s) in which the organisation/project activities are related to and one or more keywords increase its visibility and findability. For big organisations with multiple divisions (e. g., academic institutions with schools, faculties, departments, or multinational

[15] A resource can consist of one or more media parts, which must be described separately, for example, for a corpus of video recordings and their subtitles in various languages, the language value must be indicated separately for each part.

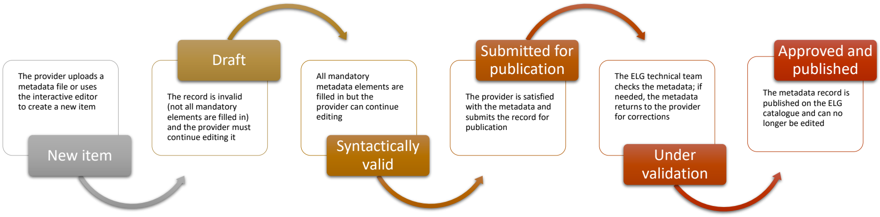
[16] https://www.rfc-editor.org/info/bcp47

companies with branches), both the parent organisation and division(s) can be registered and a link between them added.

For standardisation purposes, the ELG schema favours controlled vocabularies over free-text fields, especially when these are associated with internationally acknowledged standards, best practices or widespread vocabularies, e. g., ISO 3166 for region codes (ISO 2020), RFC 5646 for languages[17] (Phillips and Davis 2009), etc. The implementation in the form of an XML Schema Definition (XSD) imports elements from two ontologies, i. e., the MS-OWL ontology, which includes most elements and controlled vocabularies, and the OMTD-SHARE ontology[18] (Labropoulou et al. 2018) reserved for the controlled vocabularies of LT categories (also referred to as "LT taxonomy"), data formats, annotation types and methods.

6 Publication Life Cycle

ELG considers the quality of metadata records to be of primary importance as it contributes to the discovery and usage of resources. We defined a set of policies that take into account the source and the process through which a record has been entered in the ELG catalogue.



The provider uploads a metadata file or uses the interactive editor to create a new item

The record is invalid (not all mandatory elements are filled in) and the provider must continue editing it

All mandatory metadata elements are filled in but the provider can continue editing

The provider is satisfied with the metadata and submits the record for publication

The ELG technical team checks the metadata; if needed, the metadata returns to the provider for corrections

The metadata record is published on the ELG catalogue and can no longer be edited

Fig. 5 ELG publication life cycle

The ELG publication life cycle consists of a set of states through which an entry progresses, from its creation in the ELG platform until it is published (Figure 5). A *new item* is created each time a provider adds a new metadata record. The record can remain at the *draft* status as long as the provider wishes, in which case no validation checks are made – apart from validation of the data types of the metadata elements (e. g., that a URL is properly formulated). At the *syntactically valid* status, a metadata record must comply with the minimal version of the ELG schema (i. e., all mandatory elements must be filled in). The provider can still continue to edit it until they are satisfied with the description and can then submit it for publication; once submitted, the provider is notified by email. While the record is *submitted for publication* the

[17] https://datatracker.ietf.org/doc/html/rfc5646

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

entry is validated at the metadata, technical and legal level. The validation, which is described in more detail in Chapter 3, aims to check the consistency of the description and, where required, its technical compliance with the ELG specifications; it does not include any qualitative evaluation of the resource itself. The validation is currently performed by the ELG team. When validators identify a problem, they contact the provider and recommend changes and additions to the metadata; in such cases, the status is changed to *syntactically valid* again and the provider is notified to make the appropriate amendments. When the validators have approved an item, it is automatically visible via the ELG catalogue. Published metadata records cannot be edited any more, i. e., they are immutable.

Metadata records added by individuals go through the whole publication life cycle. Human validation aims at ensuring a minimum level of quality included in the records, which can be achieved through interactions with the provider. This procedure cannot be adopted for metadata records automatically imported from other catalogues. For these, the responsibility for the quality and extent of information lies with the source catalogue. The same policy, that of accepting records as is, has been adopted for records added through bulk initiatives, such as the collaborative survey of LRTs undertaken in the context of the European Language Equality project[19] and described in Chapter 6.

7 ELG and the FAIR Principles

The publication of the FAIR principles (Wilkinson et al. 2016) marked a landmark for infrastructures that support the sharing and re-use of data resources. The FAIR principles are guidelines set to enhance re-usability of data by improving their findability, accessibility, interoperability and re-usability. They are intended both for humans and machines, and put an emphasis on machine actionability, i. e., the capacity of computational systems to find, access, interoperate, and reuse data with no or minimal human intervention.[20] ELG has implemented mechanisms and policies to ensure that resources (data and software) included in ELG as well as the metadata that describe them are FAIR, i. e., adhere to the FAIR principles.[21]

Findability principles

- *F1 – (Meta)data are assigned a globally unique and persistent identifier*
  Resources hosted in ELG and ELG-compatible services are assigned a DOI (Digital Object Identifier)[22] provided by DataCite[23]. Metadata for resources will also have their own unique identifier created on the basis of the resource

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

[20] https://www.go-fair.org/fair-principles/

[21] https://force11.org/info/the-fair-data-principles/

[22] https://www.doi.org

[23] https://datacite.org

DOI. For metadata records that do not have an accompanying file and hence cannot be assigned a DOI, we use their URL as an identifier.

- *F2 – Data are described with rich metadata*
  The ELG metadata schema is rich in information. Providers are encouraged to add not only the mandatory but also recommended information. The validation process for resources and services aims at improving metadata quality.
- *F3 – Metadata clearly and explicitly include the identifier of the data they describe*
  The element “identifier” (with the “identifier scheme” attribute) is included in the metadata record.
- *F4 – (Meta)data are registered or indexed in a searchable resource*
  All metadata records are indexed and searchable in the ELG catalogue and also accessible to search engines. In addition, we expose the metadata records of LRTs to Google’s dedicated search engine for research datasets.[24]

Accessibility principles

- *A1 – (Meta)data are retrievable by their identifier using a standardised communications protocol*
  All metadata in ELG are accessible via the ELG catalogue. Resources hosted in ELG and ELG-compatible are accessible via their DOI and directly retrievable via a URL. The HTTPS protocol is used.
- *A1.1 The protocol is open, free, and universally implementable*
  HTTPS is used for providing access to metadata and resources.
- *A1.2 The protocol allows for an authentication and authorisation procedure, where necessary*
  HTTPS is used for providing access to metadata and resources. ELG uses an authentication and authorisation system.
- *A2 – Metadata are accessible, even when the data are no longer available*
  When a resource or a metadata record is deleted, a tombstone page with all the required elements following DataCite recommendations is put in place.

Interoperability principles

- *I1 – (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation*
  All metadata records are exported in XML format, a subset is available in JSON-LD format; work is ongoing for the export into RDF using the MS-OWL ontology.
- *I2 – (Meta)data use vocabularies that follow FAIR principles*
  The metadata elements and values are taken from two RDF/OWL ontologies, MS-OWL and OMTD-SHARE[25].

[24] https://datasetsearch.research.google.com

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

- *I3 – (Meta)data include qualified references to other (meta)data*
  Qualified relations are used for linking between versions of the resources and, in cases of imported records, for linking with their source metadata records.

Re-usability principles

- *R1 – (Meta)data are richly described with a plurality of accurate and relevant attributes*
  Alongside the "description" element where providers are advised to add as much information as possible for the benefit of human users, the ELG schema includes elements that can be used to identify potential uses of a resource and properties that make clear where they can be of use, e. g., "intended application", "service function", "domain", etc.
- *R1.1 – (Meta)data are released with a clear and accessible data usage license*
  All resources must have a licence; the licence value and a link to the licence text are included in the metadata. Metadata are also permissively licensed with a Creative Commons licence.
- *R1.2 – (Meta)data are associated with detailed provenance*
  The source for the metadata record is explicitly added in the metadata record ("metadata creator" or "source repository"). Properties about the creation of a resource are included in the metadata.
- *R1.3 – (Meta)data meet domain-relevant community standards*
  With regard to the metadata, the ELG schema is based on META-SHARE, a well-established metadata vocabulary in the LT community. For the tools and services added in the ELG catalogue, the technical specifications follow current best practices (e. g., preparing a Docker image). For data, a set of recommendations, taking into account established file formats, standards, and de facto best practices, is under construction.

8 Related Platforms and Infrastructures

ELG builds upon previous work of the ELG consortium partners and the wider European LT community (Rehm et al. 2020b), especially META-NET[26] and ELRC[27].

The ELG platform shares common features and goals with other platforms, repositories, projects or other initiatives: 1. a collection of LT/NLP tools or datasets, 2. a platform, which harvests metadata records from distributed sources, 3. a platform for the sharing of tools or datasets, 4. a platform for the deployment of services, 5. a repository for storing data files. Comparisons can be made along various dimensions. We include here an overview at the level of the main functionalities provided, while the respective background and technical details are presented in Chapters 3 and 4. An alternative and minimally outdated comparison is provided in Rehm et al. (2020a).

[26] http://www.meta-net.eu

[27] https://www.elrc-share.eu

META-SHARE[28] is a network of repositories (Piperidis 2012; Piperidis et al. 2014). Each repository, or node, hosts various types of resources (datasets, services, etc.) described with the META-SHARE metadata schema (Gavrilidou et al. 2012). Each node is deployed at a different organisation. The nodes periodically harvest metadata records from each other. Architecture and conceptual design of the ELG platform have been inspired by the META-SHARE setup but designed and implemented from scratch. ELG adopts a different approach as it operates as a centralised platform where individuals can directly register, download and run resources and services. Harvesting is also performed but from external catalogues (e. g., ELRC-SHARE[29], LINDAT/CLARIAH-CZ[30], etc.), as described in Chapter 6. From an engineering point of view, ELG is a radically improved version of META-SHARE, e. g., 1. ELG offers REST APIs while META-SHARE does not, 2. the ELG front end and back end are implemented as different layers that can be developed in parallel, 3. the metadata schema has been updated and extended to cover new resource types and description requirements.

The OpenMinTeD platform[31] was designed as an open, service-oriented e-Infrastructure for Text and Data Mining of scientific content (Labropoulou et al. 2018). It includes a catalogue for datasets, NLP and text mining services, worfklows, lexica etc., described with a rich metadata schema, OMTD-SHARE. REST APIs for searching, metadata and resource upload/download are provided, as in the case of ELG. OpenMinTeD was a centralised repository, and harvesting was employed as a one-off procedure for importing metadata records from a few content providers. It supported the creation of workflows from tools contained in the catalogue, and their execution on datasets provided through the same platform; the functionality was based on the Galaxy[32] worfklow management system (Afgan et al. 2018).

ELRC-SHARE[33] (Piperidis et al. 2018a) is an infrastructure developed by the European Language Resource Coordination action[34] with the objective to host, document, manage and distribute LRs pertinent to MT, with a particular focus on the needs of the eTranslation[35] service of the European Commission. It is a centralised repository with a catalogue of datasets, which are added and documented by individuals. Metadata records of tools and services are listed as for information only.

The European AI-on-demand platform, as initiated by the EU project AI4EU seeks to bring together the European AI community while promoting European values.[36] The platform is a facilitator of knowledge transfer from research to multiple

[28] http://www.meta-share.org

[29] https://www.elrc-share.eu

[30] https://lindat.mff.cuni.cz

[31] https://github.com/openminted – the OpenMinTeD platform is not available online any more.

[32] https://galaxyproject.org/learn/advanced-workflow/

[33] https://www.elrc-share.eu

[34] https://lr-coordination.eu

[35] https://cor.europa.eu/en/engage/Pages/e-translation.aspx

[36] https://www.ai4europe.eu

business and industry domains. The AI catalogue[37] is designed for hosting datasets and services in the area of AI; for instance, it includes NLP resources, computer vision services, etc. The capabilities of the metadata schema used are rather limited compared to the ELG schema. It also provides catalogues for organisations involved in AI[38], collaborating projects[39] and educational resources[40], but the catalogues are all separate, without any linking between the entities as offered in the ELG catalogue.

CLARIN[41] (Hinrichs and Krauwer 2014; Eskevich et al. 2020) is a European Research Infrastructure providing access to digital language resources and tools to researchers in the humanities and social sciences. CLARIN does not host a single repository; instead, it is organised in the form of a network of centres that operate their own repositories and catalogues. The individual centres are free in their choice of repository software and metadata schema (Broeder et al. 2008). The CLARIN Virtual Language Observatory[42] is the central catalogue which harvests metadata from all centres as well as other catalogues of interest to scholars in the target disciplines and displays them in a uniform way, although only a subset of the metadata elements are common. Processing services are catalogued centrally in the Language Switchboard [43], while some CLARIN centres make available processing services connected to their catalogues or offered separately (e. g., LINDAT/CLARIAH-CZ[44], PORTULAN-CLARIN[45], CLARIN:EL[46], etc.). Unlike ELG, there is no central compute infrastructure for deploying and running processing services.

The Language Application Grid (LAPPS Grid)[47] (Ide et al. 2014, 2016) is an open, interoperable web service platform for NLP research and development. It provides facilities for selecting and combining NLP tools and services to create workflows, composite services, and applications, and to evaluate, reproduce, and share them. It is based largely on the Galaxy[48] worfklow management system and does not actually include a catalogue. Some limited metadata have to be provided in order to create the files that are required for adding tools used in Galaxy wokflows, e. g., the name of the tool, a description, input parameters etc. For datasets no metadata are required since they are not permanently stored in Galaxy.

Hugging Face[49] is an AI/NLP company, offering repository and deployment functionalities for machine learning (Wolf et al. 2020). It hosts a large set of models and

[37] https://www.ai4europe.eu/research/ai-catalog

[38] https://www.ai4europe.eu/ai-community/organizations

[39] https://www.ai4europe.eu/ai-community/projects

[40] https://www.ai4europe.eu/education/education-catalog

[41] https://www.clarin.eu

[42] https://vlo.clarin.eu

[43] https://switchboard.clarin.eu

[44] https://lindat.mff.cuni.cz

[45] https://portulanclarin.net

[46] https://inventory.clarin.gr

[47] https://www.lappsgrid.org

[48] https://galaxyproject.org/learn/advanced-workflow/

[49] https://HuggingFace.co

datasets that can be used for model training. It offers a catalogue with a limited REST API, e. g., the API does not allow filtering search results, etc. Similar to this, there are other catalogues and repositories, such as Kaggle[50] and Papers With Code[51], which target the machine learning community. These are also community-driven, i. e., resources are registered by individuals and have their own metadata schemas.

Finally, we should mention the long lasting initiative of ELRA and the LREC community in establishing the LREC Map (Calzolari et al. 2010), as well as the growing popularity of initiatives that include general (e. g., European Open Science Cloud[52]) or federated catalogues (e. g., Gaia-X[53]) and also general repositories (e. g., Zenodo[54]), which bring together a large range of resources from and for various disciplines. See Chapter 6 for more details.

9 Conclusions

ELG has been designed as the primary platform for the European LT community, adopting a holistic view of technology development, deployment and use, bringing together language data, resources and processing services as well as the commercial and non-commercial LT actors and initiatives. ELG has established and implemented a standardised resource life cycle catering for all stages, from creation to publication and version evolution. The primary services offered are dedicated to the deposition, discovery, distribution and deployment of language resources and technologies through appropriate interfaces for technical and non-technical providers, developers, consumers and integrators. Such interfaces include web GUIs, REST APIs and a Python Software Development Kit (SDK). Its operations are supported by a metadata model underlying the description, search, discovery and distribution of resources and services, conforming to the FAIR principles. On this basis, ELG has started building bridges to existing initiatives for harvesting and importing information and resources from other infrastructures, platforms and repositories under mutually agreed conditions, business policies, acknowledgement and attribution of the source, and collaborates in joint initiatives and crowdsourcing campaigns.

References

Afgan, Enis, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James

[50] https://www.kaggle.com

[51] https://paperswithcode.com

[52] https://eosc-portal.eu

[53] https://www.gaia-x.eu

[54] https://zenodo.org

Taylor, Anton Nekrutenko, and Daniel Blankenberg (2018). “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update”. In: *Nucleic Acids Research* 46.W1, W537–W544. DOI: 10.1093/nar/gky379. URL: https://academic.oup.com/nar/article/46/W1/W537/5001157.

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez-Beltran, Andrea Perego, and Peter Winstanley, eds. (2020). *Data Catalog Vocabulary (DCAT) – Version 2*. W3C Recommendation. URL: https://www.w3.org/TR/vocab-dcat-2/.

Broeder, Daan, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, and Peter Wittenburg (2008). “Foundation of a Component-based Flexible Registry for Language Resources and Technology”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/364_paper.pdf.

Calzolari, Nicoletta, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis (2010). “The LREC Map of Language Resources and Technologies”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/370_Paper.pdf.

Eskevich, Maria, Franciska de Jong, Alexander König, Darja Fišer, Dieter Van Uytvanck, Tero Aalto, Lars Borin, Olga Gerassimenko, Jan Hajic, Henk van den Heuvel, Neeme Kahusk, Krista Liin, Martin Matthiesen, Stelios Piperidis, and Kadri Vider (2020). “CLARIN: Distributed Language Resources and Technology in a European Infrastructure”. In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France: ELRA, pp. 28–34. URL: https://aclanthology.org/2020.iwltp-1.5.

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.

Hinrichs, Erhard and Steven Krauwer (2014). “The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: ELRA, pp. 1525–1531. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

Ide, Nancy, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright (2016). “The Language Application Grid”. In: *Worldwide Language Service Infrastructure*. Ed. by Yohei Murakami and Donghui Lin. Cham: Springer, pp. 51–70. DOI: 10.1007/978-3-319-31468-6_4.

Ide, Nancy, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright (2014). “The Language Application Grid”. In: *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/926_Paper.pdf.

ISO (2020). *ISO 3166 – Country Codes*. International Organization for Standardization. URL: https://www.iso.org/iso-3166-country-codes.html.

Khan, Anas Fahad, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, and John P. McCrae (2022). “When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data”. In: *Semantic Web Journal*. Accepted for publication.

Labropoulou, Penny, Dimitris Galanis, Antonis Lempesis, Mark Greenwood, Petr Knoth, Richard Eckart de Castilho, Stavros Sachtouris, Byron Georgantopoulos, Stefania Martziou, Lucas Anastasiou, Katerina Gkirtzou, Natalia Manola, and Stelios Piperidis (2018). “OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content”. In: *Proceedings of WOSP 2018 (co-located with LREC 2018)*. Miyazaki, Japan: ELRA, pp. 7–12. URL: http://lrec-conf.org/workshops/lrec2018/W24/pdf/13_W24.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/.

Maali, Fadi and John Erickson, eds. (2014). *Data Catalog Vocabulary (DCAT) – Version 1*. W3C Recommendation. URL: https://www.w3.org/TR/2020/SPSD-vocab-dcat-20200204/.

McCrae, John Philip, Penny Labropoulou, Jorge Gracia, Marta Villegas, Víctor Rodríguez-Doncel, and Philipp Cimiano (2015). "One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web". In: *The Semantic Web: ESWC 2015 Satellite Events*. Ed. by Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann. Lecture Notes in Computer Science. Springer International Publishing, pp. 271–282. URL: https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42.

Phillips, Addison and Mark Davis (2009). *Tags for Identifying Languages*. Tech. rep. RFC 5646. Internet Engineering Task Force. URL: https://datatracker.ietf.org/doc/rfc5646.

Piperidis, Stelios (2012). "The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: ELRA.

Piperidis, Stelios, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou (2018a). "Managing Public Sector Data for Multilingual Applications Development". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/648.pdf.

Piperidis, Stelios, Penny Labropoulou, and Maria Gavriilidou (2018b). "clarin:el An infrastructure for the documentation, sharing and processing of language data (in Greek)". In: *Proceedings of the 12th International Conference on Greek Linguistics (ICGL12)*. Vol. 2. Berlin, Germany: Edition Romiosini/CeMoG, Freie Universität Berlin, pp. 851–869. URL: http://www.cemog.fu-berlin.de/en/icgl12/offprints/piperidis-lampropoulou-gavriilidou/icgl12_Piperidis-et-al.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, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch (2020a). "European Language Grid: An Overview". 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. 3359–3373. URL: https://www.aclweb.org/anthology/2020.lrec-1.413/.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020b). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". 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. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

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.

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.

Appendix

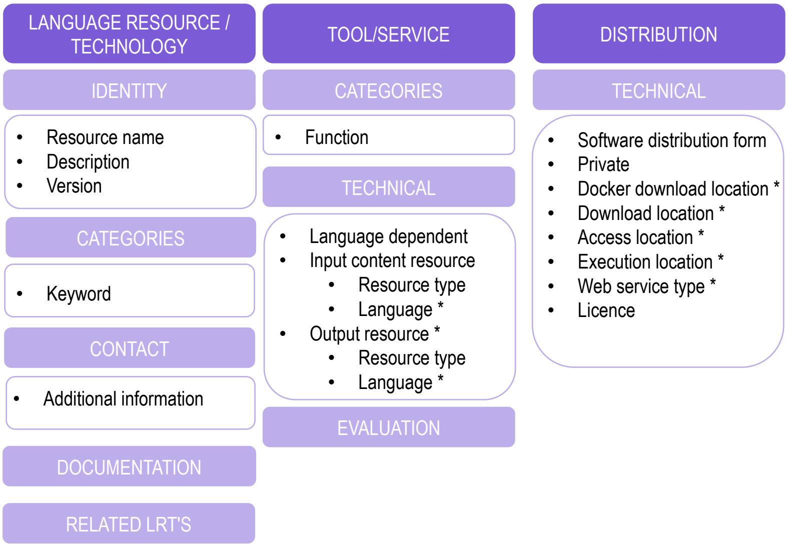


Fig. 6 ELG minimal schema version for a tool/service

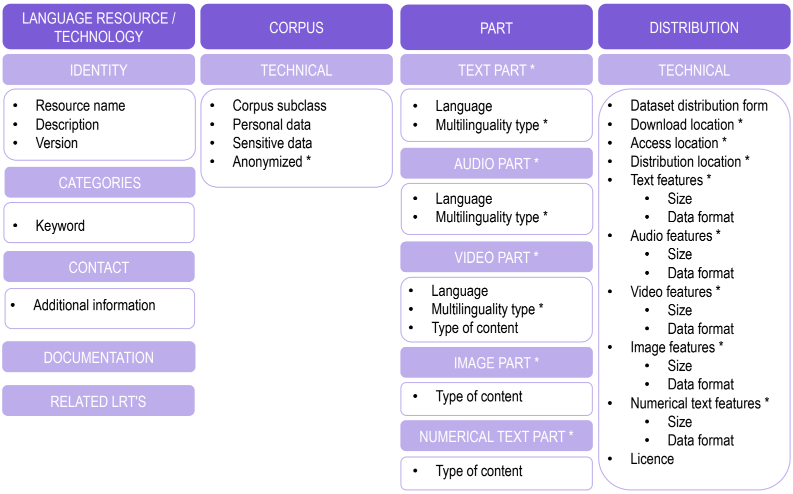


Fig. 7 ELG minimal schema version for a corpus

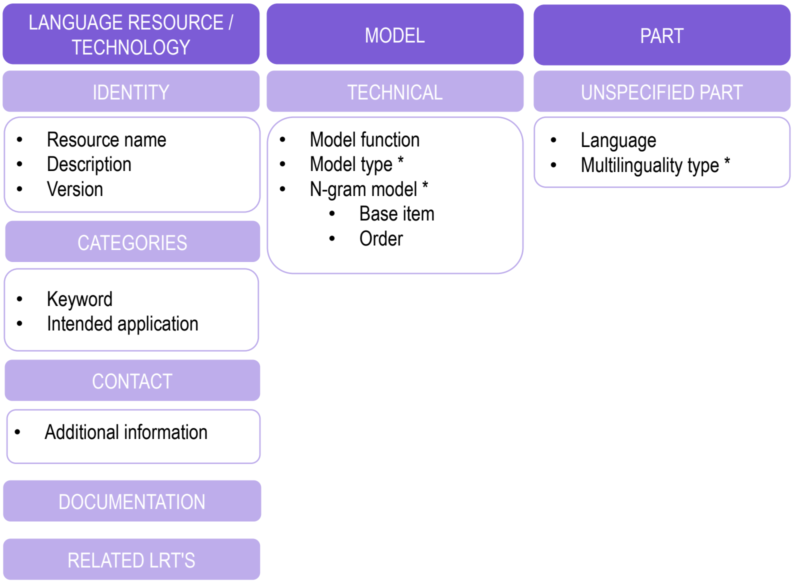


Fig. 8 ELG minimal schema version for a model

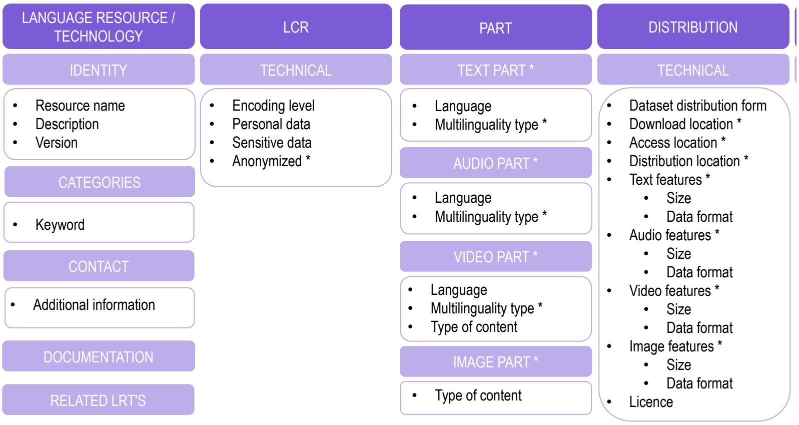


Fig. 9 ELG minimal schema version for a lexical/conceptual resource

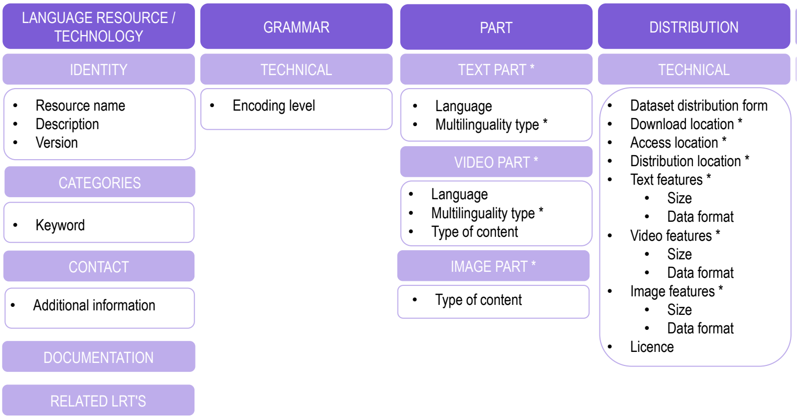


Fig. 10 ELG minimal schema version for a grammar