
European Language Grid

Release 1

Oct 21, 2020

1	Overview	3
2	Browse, search & view	5
2.1	View catalogue	5
2.2	View catalogue entry	5
3	Download a resource	15
4	Test an LT service	19
5	Provide Language Resources and Technologies	23
5.1	Minimal requirements	23
6	Provide a functional LT service	33
6.1	How an LT Service is integrated to ELG	33
6.2	Technical Requirements	35
6.3	Describe a functional LT service	36
6.4	Register an LT Service to the platform	49
6.5	Frequently Asked Questions	51
7	Provide a Language Resource	55
7.1	Technical requirements	56
7.2	Describe a Language Resource	56
7.3	Register a language resource to the platform	56
8	Describe a corpus (dataset)	59
8.1	Examples of metadata records for corpora	59
8.2	Minimal version metadata for corpora	67
9	Describe a language description (model, grammar)	83
9.1	Examples of metadata records for language descriptions	83
9.2	Minimal version metadata for language descriptions	89
10	Describe a lexical/conceptual resource (lexicon, glossary, ontology, etc.)	97
10.1	Examples of metadata records for lexical/conceptual resources	97
10.2	Minimal version metadata for lexical/conceptual resources	107
11	Describe a project	115

11.1	Examples of metadata records for projects	115
11.2	Minimal version metadata for projects	121
12	Describe an organization	129
12.1	Examples of metadata records for organizations	129
12.2	Minimal version metadata for organizations	134
13	Internal LT Service API specification	143
13.1	Basic API pattern	144
13.2	Utility datatypes	144
13.3	Request structure	145
13.4	Response structure	149
13.5	Progress Reporting	152
13.6	Appendix: Standard status message codes	154
14	Public LT API specification	155
14.1	Input formats	155
14.2	Service responses	156
14.3	Asynchronous processing	157
15	Dockerization	159
15.1	Build/Store Docker images	159
15.2	Dockerization of a Python-based LT service/tool	160
15.3	Dockerization of a Java-based tool	161
16	ELG metadata schema	163
16.1	Basic notions	163
16.2	Full schema documentation	165
16.3	Minimal version	165
16.4	Template - Explanations	165
17	Indices and tables	167
	Index	169

Welcome to the European Language Grid User Manual.

The [European Language Grid \(ELG\) platform](#) offers access to a **multitude of assets related to Language Technology (LT)**, including *commercial* and *non-commercial Language Technologies* for all European languages, *data resources* (such as models, datasets, lexica, terminologies, grammars), as well as information on *LT-related projects, organizations, and groups*.

This manual aims to guide

- **consumers** to exploit the ELG platform in order to find LT services, applications and resources, but also find relevant information such as companies, academic organisations and individual researchers active in LT and connect with them
- **providers** of Language Resources and Technologies to enrich the content of the platform.

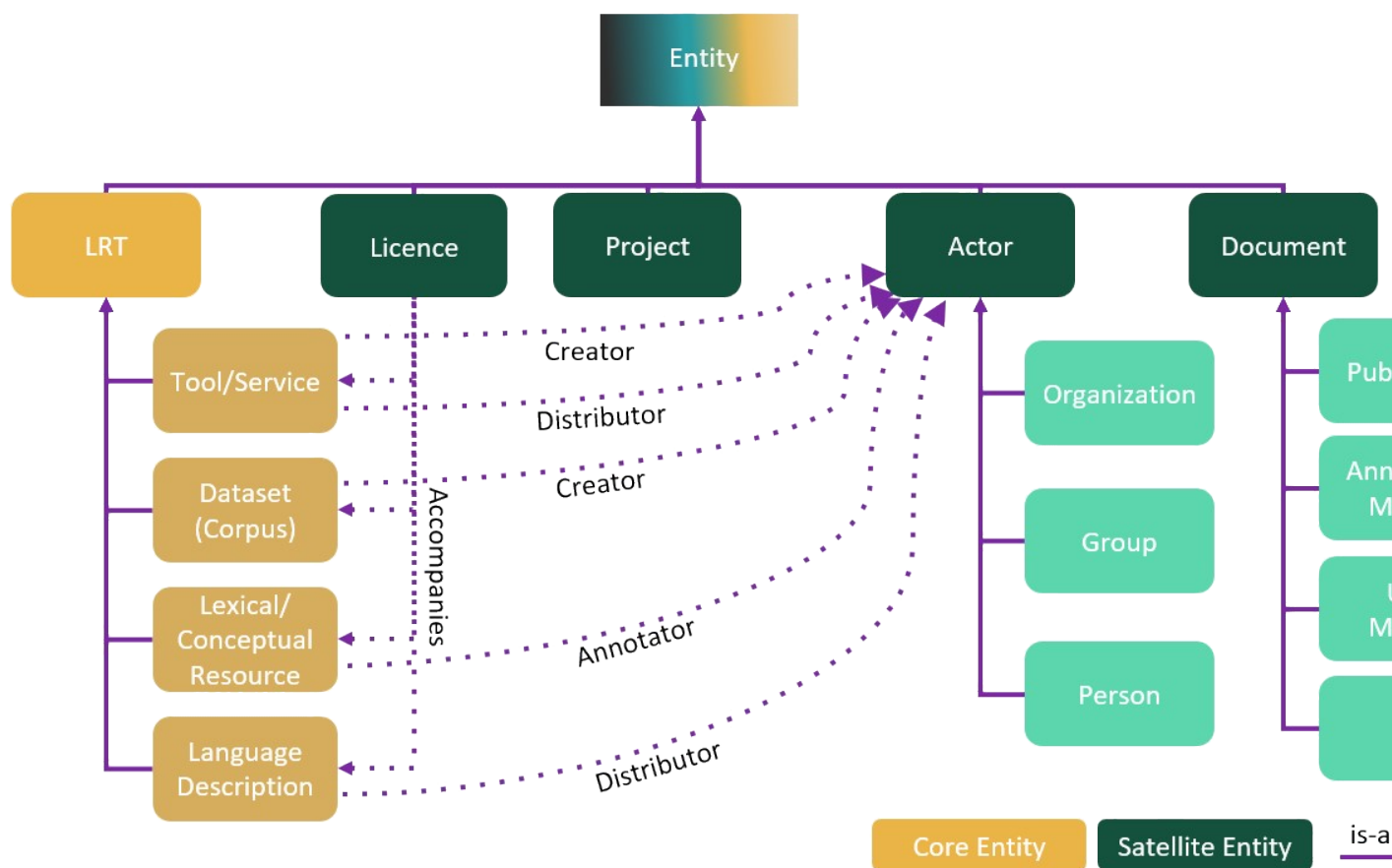
The current version of the manual documents the first official release of the ELG platform, launched in May 2020, which comes with a limited set of functionalities. More functionalities will be added at later stages, and this manual will keep on being updated following the evolution of the ELG platform.

You can visit the [ELG Catalogue](#) and browse through the content. Please, note that some functionalities are restricted to registered users.

If you have any questions or want to share feedback, please send an email to contact@european-language-grid.eu.

The [ELG catalogue](#) includes:

- **Language Resources and Technologies (LRTs)**, further classified into:
 - **tools & services**: *functional services* fully integrated and deployable in the ELG platform, but also downloadable *tools*, software in the form of source code, etc.,
 - **corpora**: collections of text documents, audio transcripts, audio and video recordings, etc.,
 - models & computational grammars, collectively referred to as **language descriptions**,
 - **lexical/conceptual resources**, which comprise computational lexica, gazetteers, ontologies, term lists, etc.
- related activities and stakeholders from the wider area of Language Technology:
 - **projects** that have funded the development of LRTs or in which they have been deployed,
 - **organizations**, as well as **groups** and **persons** active in Language Technology in Europe.



Note: The current release doesn't display *groups* and *persons*. Moreover, *documents* and *licences* are described as separate entities, but they are not shown as main entities in the catalogue.

Note: For the purposes of this document, we use the term “Language Technologies” as equivalent to “tools and services”; “Language Resources” or “Data Resources” is used for corpora, language descriptions, and lexical/conceptual resources.

Browse, search & view

2.1 View catalogue

You can

- browse through the catalogue and see all the entries (currently sorted by type and in alphabetical order),
- search for specific entries using the free text bar (e.g., you can search for a language processing service or application, a corpus, a project or an organization by its name; or search for all entries that mention “machine translation” or “English”),
- filter the catalogue or refine your search results by language, service function, resource type, etc., through the facets on the left side bar.

By clicking on the name of an entry, you can view its detailed description.

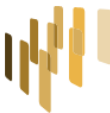
2.2 View catalogue entry

For each catalogue entry, we display a set of descriptive and technical information (metadata), together with hyperlinks to supporting documentation and other useful material. In addition, you can download, in accordance with licensing terms, resources that are available in a downloadable form, and try out or execute ELG-compliant services.

Note: For the current release, services can only be tested and only by registered users.

The following figure shows the catalogue entry for a **tool/service**.

- **Overview:** It contains the main descriptive and technical information (e.g., description of basic features, function, input and output language(s) and data format(s), etc.), links to supporting documentation, contact details, resource providers, etc.
- **Download/Run:** The second tab includes the licensing terms under which the tool/service can be accessed, and relevant technical information (i.e., whether it can be downloaded and executed locally, is provided with source code, etc.).



EUROPEAN
LANGUAGE
GRID

ALPHA

Upload

Technologies

Resources

Community

Events

About ELG

Search for services, tools, datasets, organizations...

Search

Clear all filters

Language resources & technologies

+ Corpus (252)

+ Tool/Service (157)

+ Lexical/Conceptual resource (21)

+ Language description (7)

Languages

+ English (300)

+ Polish (64)

+ German (54)

+ French (52)

+ Modern Greek (1453-) (44)

Show more

Service functions

+ Part-of-Speech Tagging (72)

+ Tokenization (67)

+ Dependency parsing (60)

+ Lemmatization (60)

+ Morphological annotation (60)

Show more

Licences

+ Open under PSI (90)

+ Creative Commons Attribution Non Commercial Share Alike 4.0 International (64)

+ Creative Commons Attribution 4.0 International (63)

+ Mozilla Public License 2.0 (62)

+ Public Domain (44)

Show more

Related entities

+ Project (10)

+ Organization (8)

455 search results

2006 CoNLL Shared Task - Ten Languages

2006 CoNLL Shared Task - Ten Languages consists of dependency treebanks in ten languages used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The languages covered in this release are: Bulgaria ...

Keyword: corpus

Languages: Swedish Danish Portuguese Spanish Turkish ...

Licences: ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

2007 CoNLL Shared Task - Basque, Catalan, Czech & Turkish

2007 CoNLL Shared Task - Basque, Catalan, Czech & Turkish consists of dependency treebanks in four languages used as part of the CoNLL 2007 shared task on multi-lingual dependency parsing and domain adaptation. The langu ...

Keyword: corpus

Languages: Catalan Czech Turkish Basque

Licences: ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0 ...

2007 CoNLL Shared Task - Greek, Hungarian & Italian

2007 CoNLL Shared Task - Greek, Hungarian & Italian consists of dependency treebanks in three languages used as part of the CoNLL 2007 shared task on multi-lingual dependency parsing and domain adaptation. The languages ...

Keyword: corpus

Languages: Modern Greek (1453-) Italian Hungarian

Licences: ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

BMI Brochures 2011-2015 (Processed)

English translations of German BMI brochures from the last four years, in TMX format.

6

Chapter 2. Browse, search & view


[← Back to catalogue](#)


Cogito Discover Named Entity Recognizer

ESI_NER

Version: 14.3.0

ToolService

[Overview](#)
[Download/Run](#)
[Test/Try out](#)
[Code samples](#)

Annotation of entities: People, Organizations, Places, Known concepts, Unknown concepts. And also tags: urls, mail addresses, phone numbers, addresses, dates, time, measures, money, percentage, file folder.

Keyword

multilingual English

Spanish German French

NER Named Entity Recognizer

Named Entity Recognition

Reconocimiento de entidades nombradas

Intended application

Named Entity Recognition

Resource provider

[Expert System](#)
[Website](#)

Additional info

[Landing page](#)

Contact

Gómez Pérez Jose Manuel


Input content resource

Language
English Spanish; Castilian
German French

Data format
JSON text/plain HTML XML

Processing resource type
file

[more](#)


Function

Function
Named Entity Recognition

Language dependent
true



Output resource

Language
English

Data format
JSON

Processing resource type
file

[more](#)

Documentations

Is documented by

2.2. [View catalogue entry](#) [documentation](#)

Evaluated

Evaluated: true

TRL: TRL9

Resource creator



Expert System

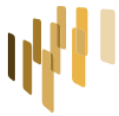
- **Try out** (only for functional services): You can provide a sample input and see the results output by the service. Depending on the type of the service, you can type in or paste some text, upload an audio file or record something, etc., and get the results rendered in a task-specific viewer.
- **Code samples** (only for functional services): You can use the code sample/template provided to test the service from the command line.

The following figure shows the entry of a **corpus**. Information is structured into tabs:

- **Overview:** It contains the main descriptive and technical information: description, subclass, keyword(s), domain(s), etc., as well as links to supporting documentation, contact details, resource providers, etc. Some properties are grouped under the “parts” of a resource, each of which is characterised by the media type (text, audio, video, image). This allows us to describe a multimedia corpus of videos, their audio excerpts (in English), the transcriptions of the recordings (in an annotated format), and the subtitles in one or more languages (English and French, provided in plain text files), as a set of four distinct parts with the corresponding properties.
- **Download:** The second tab includes the licensing terms under which the resource can be accessed, and technical details on how it can be accessed (i.e., whether it can be downloaded, used via an interface, etc.), as well as details on formats and size. If the resource has been uploaded to ELG, you will also be able to download it directly; otherwise, you will be re-directed to the original access location.
- **General:** This tab appears only for resources with a rich description and is used so as not to make the first tab too long and difficult to read.

The next two figures show the entries for **lexical/conceptual resources** (lexica, terminologies, ontologies, etc.) and **language descriptions** respectively, with information tabs similar to those of corpora.

The last two figures show respectively the catalogue entries for an **organization** and a **project**, with contact details, funding information, links to resources, etc.


[← Back to catalogue](#)


INTERA English-Slovene SVEZ ACQUIS Corpus

Version: v1.0.0 (automatically assigned)

Corpus

Overview

Download

The Slovene-English part of the INTERA corpus; written, domain specific (law) parallel subcorpus; 4MWs (2 MWs per language); TMX format.

Keyword

corpus

Intended application

machineTranslation

Domain

law

Corpus subclass

annotated corpus

Funding project

[Integrated European language data Repository Area](#)

Additional info

Contact



Gavriliidou Maria

Relations to other resources

[more](#)

Corpus parts



TEXT

 Linguality type
bilingual

 Multilinguality type
parallel

 Language
English Slovenian

 Modality type
written language

Original source description

The raw corpus comes from the SVEZ corpus provided by the Office of the Government of the Republic of Slovenia for European Affairs

Documentations

Is documented by
[Building Multilingual Terminological Resources](#)

Is documented by
[Building parallel corpora for eContent professionals](#)

Is documented by

[Language resources production models: the case of INTERA multilingual corpus and terminology](#)

Actual use

Used in application:
[terminologyExtraction](#)

Actual use details
[nlpApplications](#)


[Upload](#)
[Technologies](#)
[Resources](#)
[Community](#)
[Events](#)
[About](#)
[← Back to catalogue](#)


CEPLEXicon

Version: 1.0 (2015-04-15)


[LexicalConceptualResource](#)
[Overview](#)
[Download](#)

CEPLEXicon is a lexicon based on two different corpora of child speech – Santos corpus (Santos, 2006, Santos et al., 2014, see <http://www.clul.ul.pt/resources/546?lang=en>) and Freitas corpus (Freitas, 1997, Freitas et al. 2012). This lexicon results from the automatic tagging of the two corpora, using a tagger and the POS tag set produced in the research unit ANAGRAMA (Centro de Linguística da Universidade de Lisboa - CLUL) (Généreux, Hendrickx & Mendes, 2012). The automatic tagging was followed by a partial manual ... [Read More](#)

Keyword

lexicalconceptualresource

LCR subclass

lexicon

 Encoding level
syntax

LCR parts



TEXT

 Linguality type
monolingual

 Language
Portuguese

Funding project

[Complement Clauses in the Acquisition of Portuguese \(PTDC/CLE-LIN/120897/2015\)](#)

Additional info

[Landing page](#)

[Upload](#)[Technologies](#)[Resources](#)[Community](#)[Events](#)[About](#)[← Back to catalogue](#)

Finance English grammar

Fin.en.grm

Version: v1.0.0 (automatically assigned)

LanguageDescription

[Overview](#)[Download](#)

Finance English abnf grammar, manually created. Created within the Portdial project

Keyword

languagedescription

LD subclass

Grammar

Grammar details

Encoding level

morphology

LD parts



TEXT

Linguality type
monolingual

Language
English

Funding project

Portdial

Additional info

Landing page

Contact

Potamianos Alex


[Upload](#)
[Technologies](#)
[Resources](#)
[Community](#)
[Events](#)
[About E](#)
[← Back to catalogue](#)


Athena Research Center ARC

[Organization](#)
[Overview](#)

Athena Research and Innovation Centre (ARC) is a scientific research and technological organisation, functioning under the auspices of the General Secretariat for Research and Technology (Greek Ministry of Education). It comprises 3 research Institutes: Institute for Language and Speech Processing (ILSP), dedicated to language technology research, development and innovation; Institute for the Management of Information Systems (IMIS), dedicated to data and information management, and Industrial Systems Institute (ISI) ... [Read More](#)

LT area

[Language Technology](#)

Has division



Institute for Language and
Speech Processing
ILSP

Division category
institute

Organization description

The Institute for Language and Speech Processing (ILSP/Athena R.C.) is one of the institutes of Athena Research and Inno ... [Read More](#)

LT area

[Annotation](#)
[NLP Development Support](#)

Has division

Natural Language Processing and
Language Infrastructures
NLPLI

Division category
department

Organization description

NLPLI conducts basic and applied research in the fields of NLP and Knowledge Technologies. It designs, implements and in ... [Read More](#)

LT area

[Annotation](#)
[Machine translation support](#)
[Language resources infrastructures](#)

Organization information

ATHINA-EREVNITIKO KENTRO KAINOTOMIKIS
STIS TECHNOLOGIES TIS PLIROFORIAS
EPIKOINONION KAI TIS GNOSIS

Website

[Website](#)

Organization legal status
public organization

Organization role
LT user research organization

Address

Artemidos 6 and Epidavrou
Maroussi
Athens
151 25
GR


[Upload](#)
[Technologies](#)
[Resources](#)
[Community](#)
[Events](#)
[About](#)
[← Back to catalogue](#)


European Language Grid

Project

[Overview](#)

With 24 official EU and many more additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by thousands of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented – by nation states, languages, verticals and sectors. Likewise, while much of European LT research is world-class, with results transferr ... [Read More](#)

Keyword

[Language technology services](#)
[Multilingualism](#)
[Less-resourced languages](#)

LT area

<http://w3id.org/meta-share/omtd-share/LanguageTechnology>

Coordinator



**German Research Center for
Artificial Intelligence**

[Website](#)

Participants

Charles University

[Website](#)

SAIL LABS Technology

[Website](#)

Tilde

[Website](#)

THE UNIVERSITY OF SHEFFIELD

EVALUATIONS AND LANGUAGE RESOURCES DISTRIBUTION AGENCY

EXPERT SYSTEM IBERIA SL

2.2. View catalogue entry

Athena Research Center

[Website](#)

University of Edinburgh

[Website](#)

Project information

Website

[Website](#)

Project start date

2019-01-01

Project end date

2020-12-31

Funder

Funder



European Commission

[Website](#)

Funding scheme category

IA

Funding country

European Union

Funding type

EU funds

Grant number: 825627

Status

SIGNED

Related call

H2020-ICT-2018-2

13

Related programme

H2020

Related subprogramme

CHAPTER 3

Download a resource

You can download a resource provided in the ELG platform through the Download tab:

Downloading a resource is subject to agreeing with the licensing terms under which it is provided, while additional actions may also be required (for instance, access only for registered users):



Upload Technologies Resources Community Events About

← Back to catalogue



2006 CoNLL Shared Task - Ten Languages

Version: 1.0 (2015-12-02)

Corpus

Overview

Download

D Distribution

Dataset distribution form
downloadable

Text features

size
30 file

Data format
CoNLL-2006

Licence
ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0
http://catalogue.elra.info/static/from_media/metashare/licences/ELRA_END_USER.pdf

Membership institution
ELRA

Availability
2015-12-02 -

Distribution rights holder

 **ELRA**
 Website

Download 

D Distribution

Dataset distribution form
downloadable

Text features

size
30 file

Data format
CoNLL-2006

Licence
ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0
http://catalogue.elra.info/static/from_media/metashare/licences/ELRA_END_USER.pdf

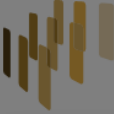
Membership institution
ELRA

Availability
2015-12-02 -

Distribution rights holder

 **ELRA**
 Website

Download 

**EUROPEAN
LANGUAGE
GRID**ALPHA


UploadTechnologiesResourcesCommunityEventsAbout E

Accept the Licence Agreement

1 of 4

Automatic Zoom

evaluations and language
resources distribution Agency



9, rue des Cordelières

75 013 Paris

Tél. : 01 43 13 33 33

Fax : 01 43 13 33 30

Web : www.elda.org

Email : info@elda.org

**LANGUAGE RESOURCES
END-USER AGREEMENT**

(Agreement Ref. No. LC/ELDA/END-USER/2015/000/NAME)

This agreement is made by and between:

Membership institution
ELRA

Availability
2015-12-02 -

Membership institution
ELRA

Availability
2015-12-02 -

Disagre

CHAPTER 4

Test an LT service

You can try out LT services that follow ELG technical specifications with a User Interface (UI) provided by the ELG platform, or via the command line. The catalogue entry of a service includes two tabs for this:

Try out: You can provide a sample input and see the results output by the service. Depending on the type of the service, you can type in or paste some text, upload an audio file or record something, etc., and get the results rendered in a task-specific viewer.

Code samples: You can use the code sample provided to test the service from your command line.

Note: For the current release, only registered users can try out services.

More information on how to call and test an integrated LT service is given in the *Public LT API specification*.



[Upload](#)

[Technologies](#)

[Resources](#)

[Community](#)

[Events](#)

[About E](#)


[← Back to catalogue](#)



ILSP Named Entity Recognizer for English

ILSP-NER-English

Version: v1.0.0

 ToolService

[Overview](#)

[Download/Run](#)

[Test/Try out](#)

[Code samples](#)

Type text to annotate

I moved to Berlin

SUBMIT



[← Back to catalogue](#)



ILSP Named Entity Recognizer for English

ILSP-NER-English

Version: v1.0.0

 ToolService

Overview

[Download/Run](#)

Test/Try out

Code samples

If you want to send a processing request from your command line to ILSP Named Entity Recognizer for English you can use the following code sample

```
curl -H "Authorization: Bearer $TOKEN" -H "Content-Type: text/plain" --data-binary "@./path/somefile" https://live.european-language
```

Token:

eYjhbGcI0iJSUzIlN1IsInR5cCIgO1AiS1SU1iWia2lK1iA6ICJ0MVhwNX1Cb1VpREM0emxOdTcyUNwT3hKRGI1fQTDObVRkSjZrCvRIQW5nIn0.eYJdgqkIOiIZZjM5MDI3Y
eHAiOjE1ODc3NDU2MzaSm5iZi16MCwiaWF0IjoXNTg3NzM0ODMwLJCjc3MiOiJodHRCwzcovL2xpdmUuZXVybzB1YW4tbGFuZS3ZhZ2UtZ3JpZC5ldS9hdXRoL3JlYWxtcy9FTmYyZGE2ZWNmLTVlYWYtNDNhYyO4NzZlLTlyMDA5ZA4MTQ2NiIsInR5cCI6IkJlYXJlc1IsImF6cCI6InJlYWNOlWNsaWVudCIsIm5vbml1IjoIQOG1OTM2YjEtYTtkOmY0O0
IjoXNTg3NzIzMdQ1LCJzZXNzaW9uX3N0YXR1Ijo1ZDJKOWUYNGYtZTE4ZC00YzQ3LWI2NTctYmY3ZhkhZmE2NDc5Iiw1YWNYIjo1MCI5ImFsbg93ZQQtbb3JpZ2lucyI6WyJwd
ZC5ldS8iXSwicmVhbGl1fWNjZXNzIjp7InJvbGVzIjpbIkdVMRYlhZG1pb1JdfSwicmVzb3VyZ2VfWNjZXNzIjpw7InJlYWNOlWNsaWVudCIEyJyb2xlcyI6WyJwc92aWR1c
cHJvZmlsZSBibWFPbCIsImVtYWlsX3Zlcm1maWVKIjpwcnVlLCJyb2xlcyI6WyJFTTEetYWRtaW4iXSwiImFtZSI6IktpbG1hbiBFdmFuZyIsInByZWZlcnJlZD91c2VybmlFTz
Ijo1ZW4iLCJkaXNzIjBlbuYWllIjo1S2lsaWVuFiWiZmFtaW5X2S25hbWUOIjFdmFuZyIsImVtYWlsIjoia2lsaWVuLmV2YW5nQGRRMa2kuZGUifQ.HQeSe6OfR2qKPYPDKkbXNgvr
OIn_rTW_SwGN8NRqd0HevG_d7dnrlXPc4fhzn5rZUrxxzwg3Pc8rRiboGeoRCzPjrOPLeVsPvQf7TjLV-
Hgrqi3rd0lbBgqNDMIbz14qVpQmbFaKy_flgQQGTGIvRky0W6Yz6YrWpa19cJ2fx0Uez1PqrKOmAZN9t7S8cjbuECLiFhNXjb9KZ64sd12dU2OGPXNAKMFTf7G4YW4h7Qey
OPijUuvjVlvIw2e3AQrz4g

Provide Language Resources and Technologies

5.1 Minimal requirements

To share a Language Resource or Language Technology tool/service through the ELG platform, you must

- register to the platform and ask (by email to contact@european-language-grid.eu) to be granted “provider” permissions;
- **describe** your LRT according to the *ELG metadata schema* (at least the minimal version) and **upload this description** to the platform, and
- **provide access** to it, as described in the respective sections for Language Technologies (*Provide a functional LT service*) and Language Resources (*Provide a Language Resource*).

5.1.1 Minimal version - List of elements common to all LRTs

The following section lists the metadata elements that are common to all LRTs of the minimal version, presented in the order they must have to comply with the schema XSD. Dedicated sections present the additional elements that are specific to each resource type.

For a quick guide to the ELG template, see *Template - Explanations*.

Note: For this release, you **MUST** provide an ELG-compliant XML file. Upcoming releases will also include a metadata editor and other functionalities supporting an easy import of metadata records.

MetadataRecord (M)

Path MetadataRecord

Data type component

Optionality Mandatory

Explanation & Instructions

A set of formalized structured information used to describe the contents, structure, function, etc. of an entity, usually according to a specific set of rules (metadata schema)

The `MetadataRecord` element includes a set of administrative data, of which the main elements that are **required** for metadata records registered by individuals are:

- `metadataCreator`: the person that has created the metadata record; you must provide at least the given name and surname of the person
- `metadataCurator`: the person that will be assigned the responsibility to update the metadata record when imported in the ELG database; you must provide at least the given name and surname of the person
- `metadataCreationDate`: the date when the metadata record was created
- `compliesWith`: for ELG metadata records, this must always be the ELG-SHARE metadata schema

Example

```
<ms:MetadataRecord>
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.org/
  ↪meta-share/meta-share/elg">default id</ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-02-28</ms:metadataCreationDate>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Brown</ms:surname>
    <ms:givenName xml:lang="en">George</ms:givenName>
  </ms:metadataCreator>
</ms:MetadataRecord>
```

resourceName (M)

Path `MetadataRecord.DescribedEntity.LanguageResource.resourceName`

Data type multilingual string

Optionality Mandatory

Explanation & Instructions

Introduces a human-readable name or title by which the resource is known

This is the “brand name” of your resource; try to use a name that is unique.

Example

```
<ms:resourceName xml:lang="en">GATE: English Named Entity Recognizer</ms:resourceName>
```

resourceShortName (R)

Path MetadataRecord.DescribedEntity.LanguageResource.resourceShortName

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces a short form (e.g., abbreviation, acronym , etc.) used to refer to a language resource

Example

```
<ms:resourceShortName xml:lang="en">annie-named-entity-recognizer</ms:resourceShortName>
```

description (M)

Path MetadataRecord.DescribedEntity.LanguageResource.description

Data type multilingual string

Optionality Mandatory

Explanation & Instructions

Introduces a short free-text account that provides information about the resource (e.g., service function, contents of a data resource, technical information , etc.)

Example

```
<ms:description xml:lang="en">Identifies names of persons, locations, organizations, as well as money amounts, time and date expressions in English texts automatically.</ms:description>
```

LRIdentifier (RA)

Path MetadataRecord.DescribedEntity.LanguageResource.LRIdentifier

Data type string

Optionality Recommended when applicable

Explanation & Instructions

A string (e.g., PID, DOI, internal to an organization , etc.) used to uniquely identify a language resource

You must also use the attribute LRIdentifierScheme to specify the identifier scheme (e.g., DOI, Handle, ...)

If the resource is already described in another resource and has a PID, please add it with the appropriate attribute.

For metadata records imported manually at ELG, you can use the value 'ELG id automatically assigned'; the value will be automatically replaced by an ELG identifier once imported in the ELG database

Example

```
<ms:LRIIdentifier ms:LRIIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">  
↔ELG id automatically assigned</ms:LRIIdentifier>
```

logo (R)

Path MetadataRecord.DescribedEntity.LanguageResource.logo

Data type URL

Optionality Recommended

Explanation & Instructions

Links to a URL with an image file containing a symbol or graphic object used to identify the entity

The logo is like a brand name for the resource; it is displayed next to the resource name in the catalogue.

Example

```
<logo>https://gate.ac.uk/plugins/gau-0.1/images/logo-gate.png</logo>
```

version (M)

Path MetadataRecord.DescribedEntity.LanguageResource.version

Data type string

Optionality Mandatory

Explanation & Instructions

Associates a language resource with a pattern that indicates its version; the recommended way is to follow the semantic versioning guidelines (<http://semver.org>) and use a numeric pattern of the form major_version.minor_version.patch

If no version is provided, the system will automatically assign the resource a 'v1.0.0 (automatically assigned)' value

Example

```
<ms:version>v8.6</ms:version>
```

additionalInfo (M)

Path MetadataRecord.DescribedEntity.LanguageResource.additionalInfo

Data type component

Optionality Mandatory

Explanation & Instructions

Introduces a point that can be used for further information (e.g. a landing page with a more detailed description of the resource or a general email that can be contacted for further queries)

It's a recommended practice to give at least a landing page (`landingPage`) or a general email addresss (`email`); if you want, you can also specify a contact person (see full schema for `contactPerson`)

Example

```
<ms:additionalInfo>
  <ms:landingPage>https://provider.example.com/product</ms:landingPage>
</ms:additionalInfo>

<ms:additionalInfo>
  <ms:email>product@example.com</ms:email>
</ms:additionalInfo>
```

keyword (M)

Path `MetadataRecord.DescribedEntity.LanguageResource.keyword`

Data type multilingual string

Optionality Mandatory

Explanation & Instructions

Introduces a word or phrase considered important for the description of a language resource, person or organization and thus used to index or classify it

You can repeat the element if you want to add more keywords. Keywords are used for discovery purposes; so, try to use words or phrases that you think users will use to find similar resources to yours.

Example

```
<ms:keyword xml:lang="en">Named entity recognition</ms:keyword>
<ms:keyword xml:lang="en">person</ms:keyword>
<ms:keyword xml:lang="en">location</ms:keyword>
<ms:keyword xml:lang="en">fake news</ms:keyword>
<ms:keyword xml:lang="en">tweets</ms:keyword>
```

domain (R)

Path `MetadataRecord.DescribedEntity.LanguageResource.domain`

Data type component

Optionality Recommended

Explanation & Instructions

Identifies the domain according to which a resource is classified

You must fill in the `CategoryLabel` element with a free text value. If you prefer to add a value from an established controlled vocabulary, you can also use the `DomainIdentifier` (with the attribute `DomainClassificationScheme` with the appropriate value).

Example

```
<ms:domain>
  <ms:categoryLabel xml:lang="en">EDUCATION & COMMUNICATIONS</
↪ms:categoryLabel>
  <ms:DomainIdentifier ms:DomainClassificationScheme="http://w3id.org/meta-
↪share/meta-share/EUROVOC">32</ms:DomainIdentifier>
</ms:domain>

<ms:domain>
  <ms:categoryLabel xml:lang="en">health</ms:categoryLabel>
</ms:domain>
```

resourceProvider (R)

Path `MetadataRecord.DescribedEntity.LanguageResource.resourceProvider`

Data type component

Optionality Recommended

Explanation & Instructions

The person/organization responsible for providing, curating, maintaining and making available (publishing) the resource

The resource provider is very similar to the publisher of scientific articles; it can be an individual or an organization.

For organizations you must add the name of the organizations (`organizationName`) and, if possible, the website.

For persons, you must add the given name and surname and, if possible, an email address or an identifier (such as [ORCID id](#)) to help uniquely identify them.

Example

```
<ms:resourceProvider>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="en">Organization</
↪ms:organizationName>
    <ms:website>https://provider.org/</ms:website>
  </ms:Organization>
</ms:resourceProvider>

<ms:resourceProvider>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
  </ms:Person>
</ms:resourceProvider>
```

publicationDate (R)

Path `MetadataRecord.DescribedEntity.LanguageResource.publicationDate`

Data type date

Optionality Recommended

Explanation & Instructions

Specifies the date when a language resource has been made available to the public

Publication date is important for citation purposes, just as for scientific articles. If this is the first time your resource is published, please use the same date as for `metadataCrationDate`. If the resource has been previously published in another repository, please add the date it was first provided there.

Example

```
<ms:publicationDate>2015-12-17</ms:publicationDate>
```

resourceCreator (R)

Path `MetadataRecord.DescribedEntity.LanguageResource.resourceCreator`

Data type component

Optionality Recommended

Explanation & Instructions

Links a resource to the person, group or organisation that has created the resource

The element is important for citation and acknowledgement purposes.

For organizations you must add the name of the organizations (`organizationName`) and, if possible, the website.

For persons, you must add the given name and surname and, if possible, an email address or an identifier (such as [ORCID id](#)) to help uniquely identify them.

Example

```
<ms:resourceCreator>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="en">example organization</
  <ms:organizationName>
    <ms:website>https://provider.org/</ms:website>
  </ms:Organization>
</ms:resourceCreator>

<ms:resourceCreator>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
  </ms:Person>
</ms:resourceCreator>
```

fundingProject (RA)

Path MetadataRecord.DescribedEntity.LanguageResource.fundingProject

Data type component

Optionality Recommended when applicable

Explanation & Instructions

Links a language resource to the project that has funded its creation, enrichment, extension , etc.

Funding information is important for acknowledgement purposes.

For projects, you must provide the name of the project (projectName) and, if possible, a website (website) and/or an identifier (ProjectIdentifier).

Example

```
<ms:fundingProject>
  <ms:projectName xml:lang="en">European Language Resource Coordination LOT3
↪</ms:projectName>
  <ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.org/meta-
↪share/meta-share/other">SMART 2015/1091 - 30-CE-0816766/00-92</ms:ProjectIdentifier>
  <ms:website>http://www.lr-coordination.eu</ms:website>
</ms:fundingProject>
```

intendedApplication (R)

Path MetadataRecord.DescribedEntity.LanguageResource.

Data type component

Optionality Recommended

Explanation & Instructions

Specifies an LT application for which the language resource has been created or for which it can be used or is recommended to be used

The element is important for discovery purposes.

You can use the element LTClassRecommended with one of the recommended values from the LT taxonomy (class 'Function' of the OMTD-SHARE ontology at <http://w3id.org/meta-share/omtd-share/>), or add a free text at the LTClassOther element.

You can repeat the element if the resource can be used for various applications. For instance, a part-of-speech tagger can be used as a component for Named entity recognition, for sentiment analysis, etc.

Example

```
<ms:intendedApplication>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪NamedEntityRecognition</ms:LTClassRecommended>
</ms:intendedApplication>

<ms:intendedApplication>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪SentimentAnalysis</ms:LTClassRecommended>
</ms:intendedApplication>
```

(continues on next page)

(continued from previous page)

```
<ms:intendedApplication>
  <ms:LTCClassOther>face recognition</ms:LTCClassRecommended>
</ms:intendedApplication>
```

isDocumentedBy (R)

Path MetadataRecord.DescribedEntity.LanguageResource.

Data type component

Optionality Recommended

Explanation & Instructions

Links a language resource to a document (e.g., research paper describing its contents or its use in a project, user manual, etc.) or any other form of documentation (e.g., a URL with support information) that is related to the resource

You can use this element to add

- supporting documentation (user manuals, training material, etc.) for the installation and use of your resource
- scientific publications that describe the resource.

If you want, you can use one of the more fine-grained relations to documents (see [full schema](#)).

You can repeat the element if you want to add more documents.

You must fill in the `title` element with the title of the document (or even an entire bibliographic record). When available, it's also recommended to add the `DocumentIdentifier` with the DOI of the document, or any other link to the document; if you do, use the attribute `DocumentIdentifierScheme` to indicate the identifier type.'

Example

```
<ms:isDocumentedBy>
  <ms:title xml:lang="en">Product User Manual</ms:title>
  <ms:DocumentIdentifier ms:DocumentIdentifierScheme="http://purl.org/spar/
↪datacite/url">https://www.company.org/product.pdf</ms:DocumentIdentifier>
</ms:isDocumentedBy>
```

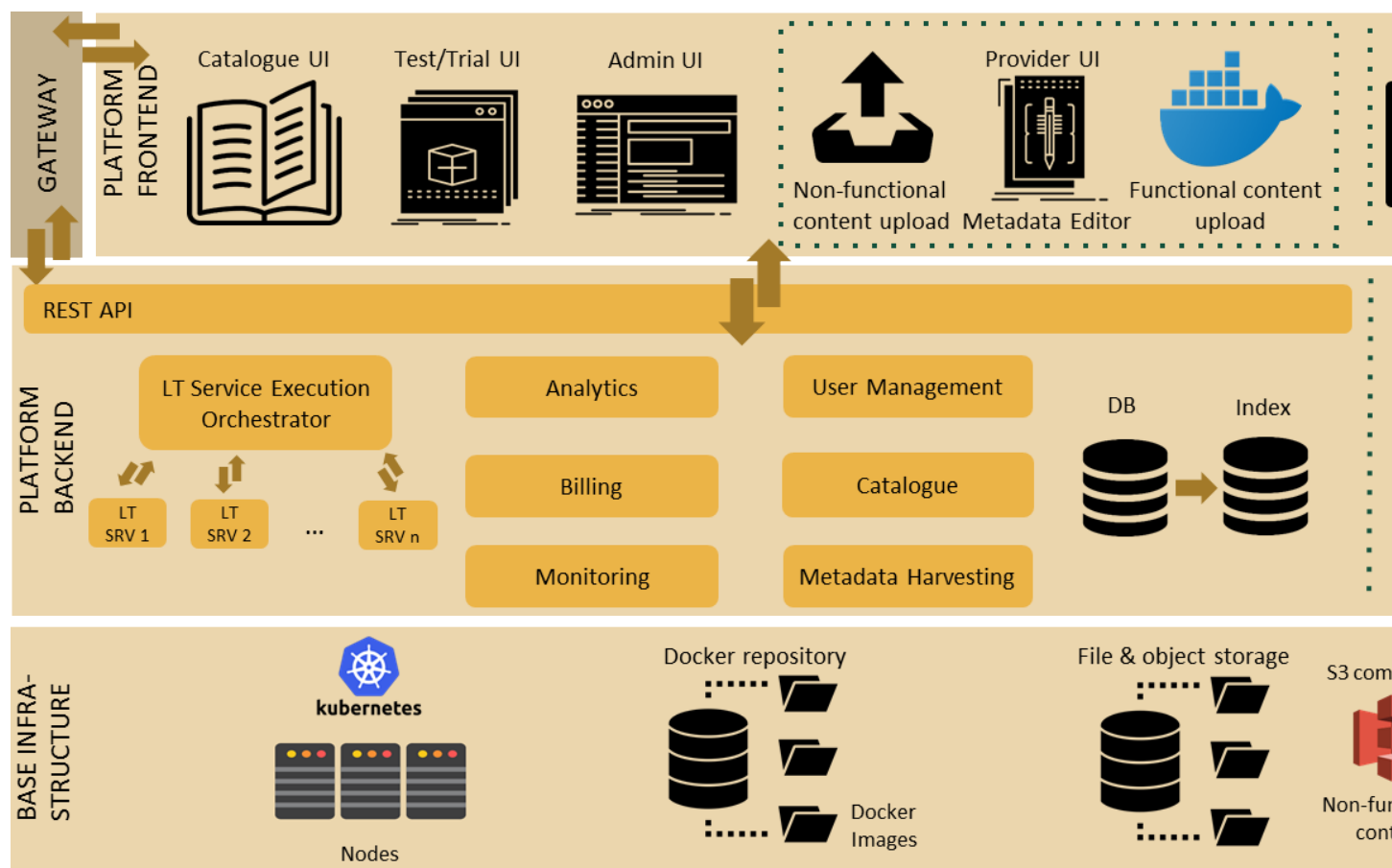
Provide a functional LT service

Currently, ELG supports the integration of tools/services that fall into one of the following broad categories:

- Information Extraction (IE),
- Text Classification (TC),
- Machine Translation (MT),
- Automatic Speech Recognition (ASR), and
- Text to Speech Generation (TTS).

6.1 How an LT Service is integrated to ELG

An overview of the ELG platform is depicted below.



The following bullets summarize how LT services are deployed and invoked in ELG.

- All LT Services (as well as all the other ELG components) are deployed (run as containers) on a Kubernetes (k8s) cluster; k8s is a system for automating deployment, scaling, and management of containerized applications.
- All LT Services are integrated into ELG via the LT Service Execution Orchestrator/Server. This server exposes a **common public REST API** used for invoking any of the deployed backend LT Services. The public API is used from ELG's Test/Trial UIs that are embedded in the ELT Catalogue; however, it can also be invoked from the command line or any programming language; see [Test an LT service](#) section for more information. Some of the HTTP endpoints that are offered in the API are given below; for more information see [Public LT API specification](#).

Endpoint	Type	Consumes	Produces
https://{domain}/execution/processText/{ltServiceID}	POST	'application/json'	'application/json'
https://{domain}/execution/processText/{ltServiceID}	POST	'text/plain' or 'text/html'	'application/json'
https://{domain}/execution/processAudio/{ltServiceID}	POST	'audio/x-wav' or 'audio/wav'	'application/json'
https://{domain}/execution/processAudio/{ltServiceID}	POST	'audio/mpeg'	'application/json'

{domain} is 'live.european-language-grid.eu' and {ltServiceID} is the ID of the backend LT service. This ID is assigned/configured during registration; see section [Register an LT Service to the platform](#) ('LT Service is deployed to ELG and configured' step).

Note: The REST API that is exposed from an LT Service X (see previous section) is for the communication between LT Service Execution Orchestrator Server and X (*ELG Internal LT API*).

- When LT Service Execution Orchestrator receives a processing request for service X, it retrieves from the database X's k8s REST endpoint and sends a request to it. This endpoint is configured/specified during the registration process; see section *Register an LT Service to the platform* ('LT Service is deployed to ELG and configured' step). When the Orchestrator gets the response from the LT Service, it returns it to the application/client that sent the initial call.

6.2 Technical Requirements

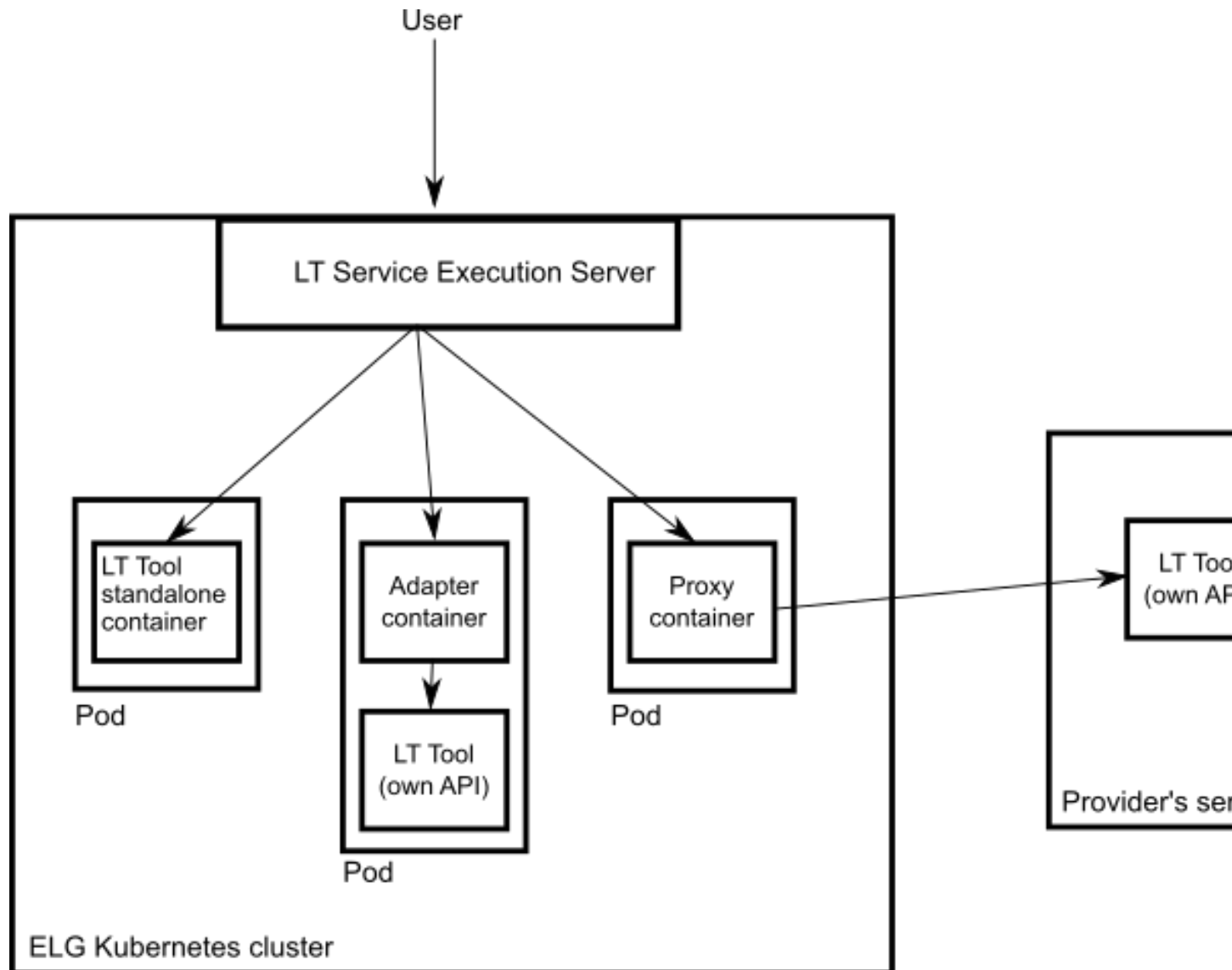
The requirements for integrating an LT tool/service are the following:

Expose an ELG compatible endpoint: You **MUST** create an application that exposes an HTTP endpoint for the provided LT tool(s). The application **MUST** consume (via the aforementioned HTTP endpoint) requests that follow the ELG JSON format, call the underlying LT tool and produce responses again in the ELG JSON format. For a detailed description of the JSON-based HTTP protocol (*ELG Internal LT API*) that you have to implement, see *the Internal LT API specification annex*.

Dockerization: You **MUST** dockerize the application and upload the respective image(s) in a Docker Registry, such as GitLab, DockerHub, Azure container registry etc. You **MAY** select out of the three following options, the one that best fits your needs:

- **LT tools packaged in one standalone image:** One docker image is created that contains the application that exposes the ELG compatible endpoint and the actual LT tool.
- **LT tools running remotely outside the ELG infrastructure:** For these tools, one *proxy* image is created that exposes one (or more) ELG compatible endpoints; the proxy container communicates with the actual LT service that runs outside the ELG infrastructure.
- **LT tools requiring an adapter:** For tools that already offer an image that exposes a non-ELG compatible endpoint (HTTP-based or other), a second *adapter* image **SHOULD** be created that exposes an ELG-compatible endpoint and acts as proxy to the container that hosts the actual LT tool.

In the following diagram the three different options for integrating a LT tool are shown:



In the [Dockerization](#) annex you will find more information on how you can create an ELG-compatible Docker image.

6.3 Describe a functional LT service

To register an ELG-compliant LT service at the platform, you must describe it according to the [ELG metadata schema](#) (at least minimal version), i.e., you have to provide a *metadata record*; some of the metadata elements are used for deploying/integrating your service to the platform.

Note: For this release, you **MUST** create an ELG-compliant XML metadata file and upload it to the platform. Upcoming releases will also provide a metadata editor as well as other functionalities supporting an easy import of metadata records.

You will find templates of metadata records for each of ELG's five categories in this [GitLab folder](#) and some examples of already registered services [here](#).

6.3.1 Examples of metadata records for LT services

ANNIE's Named Entity Recognizer (IE tool)

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../..
↳/Schema/ELG-SHARE.xsd" xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi=
↳"http://www.w3.org/2001/XMLSchema-instance">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↳org/meta-share/meta-share/elg">default id</ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-02-25</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-02-25</ms:metadataLastDateUpdated>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Roberts</ms:surname>
    <ms:givenName xml:lang="en">Ian</ms:givenName>
    <ms:email>username1@somedomain.com</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↳ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Roberts</ms:surname>
    <ms:givenName xml:lang="en">Ian</ms:givenName>
    <ms:email>username2@somedomain.com</ms:email>
  </ms:metadataCreator>
  <ms:DescribedEntity>
    <ms:LanguageResource>
      <ms:entityType>LanguageResource</ms:entityType>
      <ms:resourceName xml:lang="en">GATE: English Named Entity_
↳Recognizer</ms:resourceName>
      <ms:resourceShortName xml:lang="en">annie-named-entity-
↳recognizer</ms:resourceShortName>
      <ms:description xml:lang="en">Identify names of &lt;em>
↳persons&lt;/em>, &lt;em>locations&lt;/em>, &lt;em>organizations&lt;/em>
↳gt;, as well as &lt;em>money amounts&lt;/em>, &lt;em>time and date_
↳expressions&lt;/em> in English texts automatically. </ms:description>
      <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-
↳share/meta-share/elg">ELG id automatically assigned</ms:LRIdentifier>
      <ms:version>v8.6</ms:version>
      <ms:additionalInfo>
        <ms:landingPage>https://cloud.gate.ac.uk/shopfront/
↳displayItem/annie-named-entity-recognizer</ms:landingPage>
      </ms:additionalInfo>
      <ms:keyword xml:lang="en">Named Entity Recognition</
↳ms:keyword>
      <ms:keyword xml:lang="en">English</ms:keyword>
      <ms:resourceProvider>
        <ms:Group>
          <ms:actorType>Group</ms:actorType>
          <ms:organizationName xml:lang="en">GATE Team,
↳University of Sheffield</ms:organizationName>
          <ms:website>https://gate.ac.uk/</ms:website>
        </ms:Group>
      </ms:resourceProvider>
      <ms:publicationDate>2020-02-25</ms:publicationDate>
      <ms:resourceCreator>
        <ms:Person>
```

(continues on next page)

(continued from previous page)

```

        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Roberts</ms:surname>
        <ms:givenName xml:lang="en">Ian</ms:givenName>
        <ms:email>username3@somedomain.com</ms:email>
    </ms:Person>
</ms:resourceCreator>
<ms:intendedApplication>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/
↪ omdt-share/NamedEntityRecognition</ms:LTCClassRecommended>
</ms:intendedApplication>
<ms:LRSubclass>
    <ms:ToolService>
        <ms:lrType>ToolService</ms:lrType>
        <ms:function>
            <ms:LTCClassRecommended>http://w3id.
↪ org/meta-share/omdt-share/NamedEntityRecognition</ms:LTCClassRecommended>
        </ms:function>
        <ms:SoftwareDistribution>
            <ms:SoftwareDistributionForm>http://
↪ w3id.org/meta-share/meta-share/dockerImage</ms:SoftwareDistributionForm>
            <ms:executionLocation>http://
↪ localhost:8080/process</ms:executionLocation>
            <ms:dockerDownloadLocation>registry.
↪ gitlab.com/european-language-grid/usfd/gate-ie-tools/annie:8.6-0.0.3</
↪ ms:dockerDownloadLocation>
            <ms:licenceTerms>
                <ms:licenceTermsName xml:lang=
↪ "en">GNU Lesser General Public License v3.0 only</ms:licenceTermsName>
                <ms:licenceTermsURL>https://
↪ spdx.org/licenses/LGPL-3.0-only.html</ms:licenceTermsURL>
                <ms:LicenceIdentifier
↪ ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/SPDX">LGPL-3.0-
↪ only</ms:LicenceIdentifier>
                <ms:LicenceIdentifier
↪ ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG-ENT-LIC-
↪ 270220-00000199</ms:LicenceIdentifier>
            </ms:licenceTerms>
        </ms:SoftwareDistribution>
        <ms:languageDependent>true</
↪ ms:languageDependent>
        <ms:inputContentResource>
            <ms:processingResourceType>http://
↪ w3id.org/meta-share/meta-share/file1</ms:processingResourceType>
            <ms:language>
                <ms:languageTag>en</
↪ ms:languageTag> <ms:languageId>en</ms:languageId>
            </ms:language>
            <ms:mediaType>http://w3id.org/meta-
↪ share/meta-share/text</ms:mediaType>
            <ms:dataFormat>http://w3id.org/meta-
↪ share/omdt-share/Json</ms:dataFormat>
            <ms:characterEncoding>http://w3id.org/
↪ meta-share/meta-share/UTF-8</ms:characterEncoding>
        </ms:inputContentResource>
        <ms:outputResource>
            <ms:processingResourceType>http://
↪ w3id.org/meta-share/meta-share/file1</ms:processingResourceType>

```

(continues on next page)

(continued from previous page)

```

<ms:language>
  <ms:languageTag>en</ms:languageTag>
</ms:language>
<ms:languageTag> <ms:languageId>en</ms:languageId>
</ms:language>
<ms:mediaType>http://w3id.org/meta-
share/meta-share/text</ms:mediaType>
<ms:dataFormat>http://w3id.org/meta-
share/omtd-share/Json</ms:dataFormat>
<ms:characterEncoding>http://w3id.org/
meta-share/meta-share/UTF-8</ms:characterEncoding>
<!-- annotations: :Address, :Date,
:Location, :Organization, :Person, :Money, :Percent, :Token, :SpaceToken, :Sentence
-->
<ms:annotationType>http://w3id.org/
meta-share/omtd-share/Person</ms:annotationType>
<ms:annotationType>http://w3id.org/
meta-share/omtd-share/Location</ms:annotationType>
<ms:annotationType>http://w3id.org/
meta-share/omtd-share/Organization</ms:annotationType>
<ms:annotationType>http://w3id.org/
meta-share/omtd-share/Date</ms:annotationType>
</ms:outputResource>
<ms:trl>http://w3id.org/meta-share/meta-share/
trl4</ms:trl>
<ms:evaluated>>false</ms:evaluated>
</ms:ToolService>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

The Docker image for this LT tool is stored at GitLab registry.

Edinburgh's German to English engine (MT tool)

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi="http://
www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://w3id.org/meta-share/
meta-share/ ../../Schema/ELG-SHARE.xsd">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
org/meta-share/meta-share/elg">default id</ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-02-28</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-02-28</ms:metadataLastDateUpdated>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Germann</ms:surname>
    <ms:givenName xml:lang="en">Ulrich</ms:givenName>
    <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://w3id.org/
meta-share/meta-share/elg">ELG-ENT-PER-050320-00000787</ms:PersonalIdentifier>
    <ms:email>user@somedomain.uk</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Germann</ms:surname>
    <ms:givenName xml:lang="en">Ulrich</ms:givenName>

```

(continues on next page)

(continued from previous page)

```

    <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://w3id.org/
↪meta-share/meta-share/elg">ELG-ENT-PER-050320-00000787</ms:PersonalIdentifier>
    <ms:email>user@somedomain.uk</ms:email>
  </ms:metadataCreator>
  <ms:DescribedEntity>
    <ms:LanguageResource>
      <ms:entityType>LanguageResource</ms:entityType>
      <ms:resourceName xml:lang="en">UEDIN Machine Translation↪
↪Service for German to English</ms:resourceName>
      <ms:resourceShortName xml:lang="en">UEDIN-MT-DeEn</
↪ms:resourceShortName>
      <ms:description xml:lang="en">A machine translation (MT)↪
↪service for German-to-English translation based on the Marian machine translation↪
↪framework. The translation model is a basic transformer model trained on ca 13.3M↪
↪sentence pairs using Marian NMT</ms:description>
      <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-
↪share/meta-share/elg">ELG id automatically assigned</ms:LRIdentifier>
      <ms:version>v1.0.0</ms:version>
      <ms:additionalInfo>
        <ms:email>user@somedomain.uk</ms:email>
      </ms:additionalInfo>
      <ms:keyword xml:lang="en">Machine Translation</ms:keyword>
      <ms:keyword xml:lang="en">German</ms:keyword>
      <ms:keyword xml:lang="en">English</ms:keyword>
      <ms:keyword xml:lang="en">Neural machine translation</
↪ms:keyword>
      <ms:keyword xml:lang="en">Marian framework</ms:keyword>
      <ms:resourceProvider>
        <ms:Organization>
          <ms:actorType>Organization</ms:actorType>
          <ms:organizationName xml:lang="en">UEDIN</
↪ms:organizationName>
          <ms:OrganizationIdentifier↪
↪ms:OrganizationIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG-ENT-
↪ORG-280220-00000397</ms:OrganizationIdentifier>
          <ms:website>https://www.ed.ac.uk/informatics/
↪</ms:website>
        </ms:Organization>
      </ms:resourceProvider>
      <ms:publicationDate>2020-02-28</ms:publicationDate>
      <ms:resourceCreator>
        <ms:Person>
          <ms:actorType>Person</ms:actorType>
          <ms:surname xml:lang="en">Germann</ms:surname>
          <ms:givenName xml:lang="en">Ulrich</
↪ms:givenName>
          <ms:PersonalIdentifier↪
↪ms:PersonalIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG-ENT-PER-
↪050320-00000787</ms:PersonalIdentifier>
          <ms:email>user@somedomain.uk</ms:email>
        </ms:Person>
      </ms:resourceCreator>
      <ms:intendedApplication>
        <ms:LTClassRecommended>http://w3id.org/meta-share/
↪omtd-share/MachineTranslation</ms:LTClassRecommended>
      </ms:intendedApplication>
      <ms:LRSubclass>

```

(continues on next page)

(continued from previous page)

```

<ms:ToolService>
  <ms:lrType>ToolService</ms:lrType>
  <ms:function>
    <ms:LTClassRecommended>http://w3id.
↪org/meta-share/omtd-share/MachineTranslation</ms:LTClassRecommended>
  </ms:function>
  <ms:SoftwareDistribution>
    <ms:SoftwareDistributionForm>http://
↪w3id.org/meta-share/meta-share/dockerImage</ms:SoftwareDistributionForm>
    <ms:executionLocation>http://
↪localhost:18080/api/elg/v1</ms:executionLocation>
    <ms:dockerDownloadLocation>mt4elg-de-
↪en</ms:dockerDownloadLocation>
    <ms:additionalHWRequirements>limits_
↪memory: 2048Mi limits_cpu: 1.5</ms:additionalHWRequirements>
    <ms:licenceTerms>
      <ms:licenceTermsName xml:lang=
↪"en">CC BY-SA 4.0</ms:licenceTermsName>
      <ms:licenceTermsURL>https://
↪creativecommons.org/licenses/by-sa/4.0/</ms:licenceTermsURL>
      <ms:LicenceIdentifier_
↪ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG-ENT-LIC-
↪270220-00000097</ms:LicenceIdentifier>
    </ms:licenceTerms>
  </ms:SoftwareDistribution>
  <ms:languageDependent>true</
↪ms:languageDependent>
    <ms:inputContentResource>
      <ms:processingResourceType>http://
↪w3id.org/meta-share/meta-share/file1</ms:processingResourceType>
      <ms:language>
        <ms:languageTag>de</
↪ms:languageTag>
        <ms:languageId>de</
↪ms:languageId>
      </ms:language>
      <ms:mediaType>http://w3id.org/meta-
↪share/meta-share/text</ms:mediaType>
      <ms:dataFormat>http://w3id.org/meta-
↪share/omtd-share/Json</ms:dataFormat>
      <ms:characterEncoding>http://w3id.org/
↪meta-share/meta-share/UTF-8</ms:characterEncoding>
    </ms:inputContentResource>
    <ms:outputResource>
      <ms:processingResourceType>http://
↪w3id.org/meta-share/meta-share/file1</ms:processingResourceType>
      <ms:language>
        <ms:languageTag>en</
↪ms:languageTag>
        <ms:languageId>en</
↪ms:languageId>
      </ms:language>
      <ms:mediaType>http://w3id.org/meta-
↪share/meta-share/text</ms:mediaType>
      <ms:dataFormat>http://w3id.org/meta-
↪share/omtd-share/Json</ms:dataFormat>
      <ms:characterEncoding>http://w3id.org/
↪meta-share/meta-share/UTF-8</ms:characterEncoding>

```

(continues on next page)

(continued from previous page)

```

</ms:outputResource>
<ms:trl>http://w3id.org/meta-share/meta-share/
↪trl4</ms:trl>

<ms:evaluated>>false</ms:evaluated>
</ms:ToolService>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

The Docker image for this LT tool is stored at DockerHub.

6.3.2 Minimal version metadata

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources**, including functional LT services, are presented in *Minimal version - List of elements common to all LRTs*. **In addition**, the following metadata elements are required or recommended for tools/services:

For a quick guide to the ELG template, see *Template - Explanations*.

ToolService

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService

Data type component

Optionality Mandatory

Explanation & Instructions

Introduces the set of elements that is specific to tools/services

Example

```

<ms:LRSubclass>
  <ms:ToolService>
    <ms:lrType>toolService</ms:lrType>
    ...
  </ms:ToolService>
</ms:LRSubclass>

```

function

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.function

Data type component

Optionality Mandatory

Explanation & Instructions

Specifies the operation/function/task that a software object performs

The element is important for discovery purposes. You can fill in:

- the `LTClassRecommended` element with one of the recommended values from the [LT taxonomy](#), or
- the `LTClassOther` element with a free text.

For services that perform multiple functions (e.g., syntactic and semantic annotation) you can repeat the element.

Example

```
<ms:function>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
  ↪NamedEntityRecognition</ms:LTClassRecommended>
</ms:function>

<ms:function>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
  ↪MachineTranslation</ms:LTClassRecommended>
</ms:function>

<ms:function>
  <ms:LTClassOther>video segmentation</ms:LTClassRecommended>
</ms:function>
```

SoftwareDistribution

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.
SoftwareDistribution

Data type component

Optionality Mandatory

Explanation & Instructions

Any form with which software is distributed (e.g., web services, executable or code files, etc.)

This element groups together information that pertains to the physical form of a tool/service that is made available through the catalogue. For software that is distributed with multiple forms (e.g., as source code, as a web service, etc.), you can repeat this group of elements. The access location and the licensing conditions may differ for each distribution.

The following list includes the mandatory and recommended elements:

- `SoftwareDistributionForm` (Mandatory): The medium, delivery channel or form (e.g., source code, API, web service, etc.) through which a software object is distributed. Use the value `http://w3id.org/meta-share/meta-share/dockerImage`.
- `dockerDownloadLocation` (Mandatory if applicable): A location where the the LT tool docker image is stored. Add the location from where the ELG team can download the docker image in order to test it.
- `serviceAdapterDownloadLocation` (Mandatory if applicable): The URL where the docker image of the service adapter can be downloaded from. Required only for ELG functional services implemented with an adapter.
- `executionLocation` (Mandatory): A URL where the resource (mainly software) can be directly executed. Add here the REST endpoint at which the LT tool is exposed within the Docker image.
- `additionalHwRequirements` (Mandatory if applicable): A short text where you specify additional requirements for running the service, e.g. memory requirements, etc. The recommended format for this is: 'limits_memory: X limits_cpu: Y'

- `licenceTerms` (Mandatory): See *licenceTerms*
-

licenceTerms

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.
SoftwareDistribution.licenceTerms`

Data type component

Optionality Mandatory

Explanation & Instructions

Links the distribution (distributable form) of a language resource to the licence or terms of use/service (a specific legal document) with which it is distributed

The recommended practice is to add a licence name and identifier from the SPDX list of licences (<https://spdx.org/licenses/>). For proprietary licences or licences not included in the above list, please add a (unique) licence name and the URL where the text of the licence can be found.

Example

```
<ms:licenceTerms>
  <ms:licenceTermsName xml:lang="en">GNU Lesser General Public License v3.0 only
↪</ms:licenceTermsName>
  <ms:licenceTermsURL>https://spdx.org/licenses/LGPL-3.0-only.html</
↪ms:licenceTermsURL>
  <ms:LicenceIdentifier ms:LicenceIdentifierScheme="http://w3id.org/meta-share/
↪meta-share/SPDX">LGPL-3.0-only</ms:LicenceIdentifier>
</ms:licenceTerms>

<ms:licenceTerms>
  <ms:licenceTermsName xml:lang="en">publicDomain</ms:licenceTermsName>
  <ms:licenceTermsURL>https://elrc-share.eu/terms/publicDomain.html</
↪ms:licenceTermsURL>
</ms:licenceTerms>

<ms:licenceTerms>
  <ms:licenceTermsName xml:lang="en">Creative Commons Attribution 4.0
↪International</ms:licenceTermsName>
  <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/legalcode</
↪ms:licenceTermsURL>
  <ms:LicenceIdentifier ms:LicenceIdentifierScheme="http://w3id.org/meta-share/
↪meta-share/SPDX">CC-BY-4.0</ms:LicenceIdentifier>
</ms:licenceTerms>
```

languageDependent

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.
languageDependent`

Data type boolean

Optionality Mandatory

Explanation & Instructions

Indicates whether the operation of the tool or service is language dependent or not

For language-dependent tools/services, you will be asked to also provide the language of the input and output resources.

Example

```
<ms:languageDependent>true</ms:languageDependent>
```

inputContentResource

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.inputContentResource

Data type component

Optionality Mandatory

Explanation & Instructions

Specifies the requirements set by a tool/service for the (content) resource that it processes

The following elements are mandatory or recommended:

- `processingResourceType`: Specifies the resource type that a tool/service takes as input or produces as output; you must specify, for instance, if the tool/service can process a single file, or set of files, or processes a string typed in by the users.
- `language`: Specifies the language that is used in the resource or supported by the tool/service, expressed according to the BCP47 recommendation. See [language](#)
- `mediaType` (Recommended): Specifies the media type of the input/output of a language processing tool/service. For ELG functional services, this will be used to fit the appropriate GUI (e.g. “audio” for ASR applications, vs. “text” for Machine Translation applications)
- `dataFormat` (Recommended): Indicates the format(s) of a data resource Please, use to indicate the data format of the resource supported by the tool/service. The dataFormat CV (TODO) lists data formats, with their mimetype and documentation on the particularities, thus catering for variations of formats, e.g. GATE XML, TEI variants, etc.
- `characterEncoding` (Recommended if applicable): Specifies the character encoding used for the input/output text resource of an LT service
- `annotationType` (Recommended if applicable): Specifies the annotation type of the annotated version(s) of a resource or the annotation type a tool/ service requires or produces as an output. Use this element only if the tool/service processes pre-annotated corpora; for tools/services processing raw files, do not use. The element takes a value from a CV, see [annotationType](#).

Example

```
<!-- example for a tool with textual input -->
<ms:inputContentResource>
  <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</
  <ms:processingResourceType>
    <ms:language>
      <ms:languageTag>en</ms:languageTag> <ms:languageId>en</ms:languageId>
    </ms:language>
```

(continues on next page)

(continued from previous page)

```

    <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Json</ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
</ms:inputContentResource>

<!-- example for an Automatic Speech Recognizer -->
<ms:inputContentResource>
    <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</
↪ms:processingResourceType>
    <ms:language>
        <ms:languageTag>de</ms:languageTag> <ms:languageId>de</ms:languageId>
    </ms:language>
    <ms:mediaType>http://w3id.org/meta-share/meta-share/audio</ms:mediaType>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/mp3</ms:dataFormat>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/wav</ms:dataFormat>
</ms:inputContentResource>

```

outputResource

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.outputResource

Data type component

Optionality Recommended if applicable

Explanation & Instructions

Describes the features of the output resource processed by a tool/service.

The set of elements are the same as for the *inputContentResource*.

Make sure that you add here what is relevant for your application. For instance,

- for annotation and information extraction tools/services, use the *annotationType* to indicate the results of your processing; you can repeat it to indicate multiple annotation types (e.g., part of speech, person, amount, location, etc.)
- for Machine Translation tools, indicate the input and output languages respectively.

Example

```

<!-- example for an Information Extraction tool -->
<ms:outputResource>
    <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</
↪ms:processingResourceType>
    <ms:language>
        <ms:languageTag>en</ms:languageTag>
        <ms:languageId>en</ms:languageId>
    </ms:language>
    <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Json</ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
    <ms:annotationType>http://w3id.org/meta-share/omtd-share/Person</
↪ms:annotationType>

```

(continues on next page)

(continued from previous page)

```

    <ms:annotationType>http://w3id.org/meta-share/omtd-share/Location</
↪ms:annotationType>
    <ms:annotationType>http://w3id.org/meta-share/omtd-share/Organization</
↪ms:annotationType>
    <ms:annotationType>http://w3id.org/meta-share/omtd-share/Date</
↪ms:annotationType>
</ms:outputResource>

<!-- example for a Machine Translation tool -->
<ms:outputResource>
  <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</
↪ms:processingResourceType>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>
  </ms:language>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
  <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Json</ms:dataFormat>
  <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
</ms:outputResource>

```

language

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.
language

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Specifies the language that is used in the resource or supported by the tool/service, expressed according to the BCP47 recommendation

The element languageTag is composed of the languageId, and optionally scriptId, regionId and variantId; you can use those elements that best describe the language(s) of your resource.

Example

```

<ms:language>
  <ms:languageTag>en</ms:languageTag>
  <ms:languageId>en</ms:languageId>
</ms:language>

<ms:language>
  <ms:languageTag>en-US</ms:languageTag>
  <ms:languageId>en</ms:languageId>
  <ms:regionId>US</ms:regionId>
</ms:language>

```

implementationLanguage

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.implementationLanguage

Data type string

Optionality Recommended

Explanation & Instructions

The programming language(s) used for the development of a tool/service

Example

```
<ms:implementationLanguage>Java v8.1</ms:implementationLanguage>
```

trl

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.trl

Data type CV <http://w3id.org/meta-share/meta-share/trl>

Optionality Recommended

Explanation & Instructions

Specifies the TRL (Technology Readiness Level) of the technology according to the measurement system defined by the EC (https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf)

Example

```
<ms:trl>http://w3id.org/meta-share/meta-share/trl4</ms:trl>
```

evaluated

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.ToolService.evaluated

Data type boolean

Optionality Mandatory

Explanation & Instructions

Indicates whether the tool or service has been evaluated

If the tool/service has been evaluated, you can use the ‘evaluation’ component to give more detailed information; see [here](#) for the relevant elements.

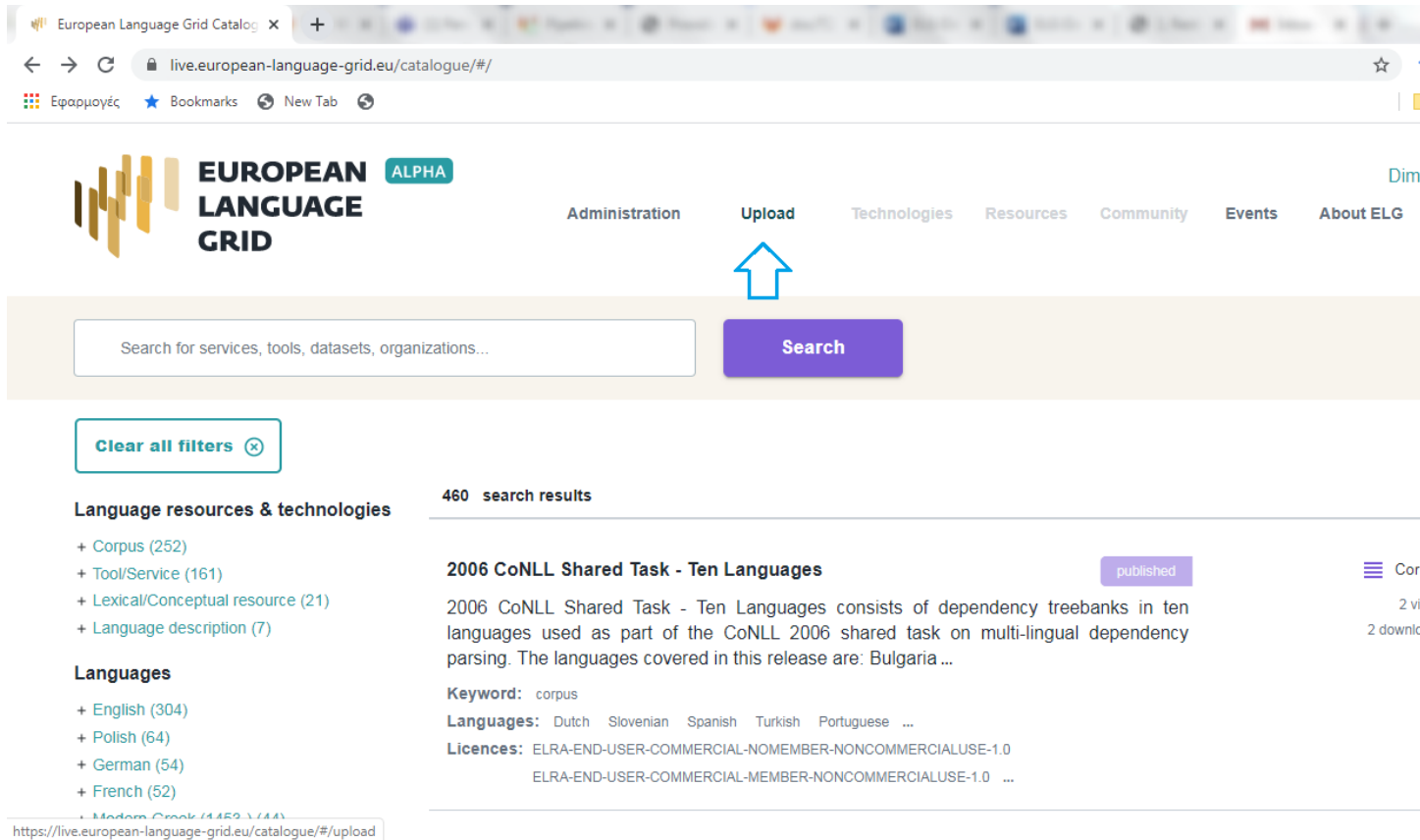
Example

```
<ms:evaluated>false</ms:evaluated>
```

6.4 Register an LT Service to the platform

The following steps should be followed:

- **Provide a metadata record:** Sign into the ELG platform using your credentials and press the “upload” button on the main menu.



Then upload the XML file that contains the metadata. In the current release, this is the only way to provide them. Upcoming releases will also include a metadata editor and other functionalities.

The metadata record is validated at import against the metadata schema. Additional rules that check for syntactic and partial semantic integrity are also used. If the file is found invalid, you will see a message with a list of errors; you must correct them and re-upload the file. If it is valid you will be shown a success message; the file will be imported in the database. At this stage, it is visible only to the platform administrators.

- **LT Service is assigned to a reviewer:** The administrator will assign it to a reviewer; during the review process, the metadata record is visible only to you (LT provider) and the reviewer.
- **LT Service is deployed to ELG and configured:** The LT service is deployed (by the reviewer) to the k8s cluster by creating the appropriate configuration yaml file and uploading to the respective GitLab repository. The CI/CD pipeline that is responsible for deployments will automatically install the new service at the k8s cluster. If you request it, a separate dedicated k8s namespace can be created for the LT service before creating the yaml file. The reviewer of the service assigns to it:

The screenshot shows the 'Upload' page of the European Language Grid (ELG) website. The browser address bar shows the URL 'live.european-language-grid.eu/catalogue/#/upload'. The page header includes the ELG logo, the word 'ALPHA', and navigation links: Administration, Upload, Technologies, Resources, Community, Events, and About ELG. Below the header, there is a 'Back to catalogue' link and a 'Upload Resources' section. A purple circle with the number '1' is positioned above the 'Upload XML' text. A large light blue box contains the text 'Drag & Drop your files or Browse'. Below this box, it says 'Upload an XML file'.

The screenshot shows the 'Search' page of the European Language Grid (ELG) website. The browser address bar shows the URL 'live.european-language-grid.eu/catalogue/#/search/geolocator'. The page header is identical to the previous screenshot. Below the header, there is a search bar containing the text 'geolocator' and a 'Search' button. A 'Clear all filters' button is also visible. The search results are displayed in a table with the following columns: Language resources & technologies, Languages, Service functions, Licences, and a search results column. The search results column shows '1 search results for geolocator'. The first result is 'geolocator', which is a 'Tool/Service' that provides 'Social media geolocation prediction for a given Tweet, or a short text.' The result is marked as 'ingested' and has a blue upward arrow icon next to it. The result also includes keywords: 'Geolocation', 'Social Media', and 'Twitter'; language: 'English'; and license: 'GNU General Public License v3.0 only'.

- the k8s REST endpoint that will be used for invoking it. The endpoint follows this template: `http://{k8s service name for the registered LT tool}.{k8s namespace for the registered LT tool}.svc.cluster.local{the path where the REST service is running at}`. The {the path where the REST service is running at} part can be found in the `executionLocation` field in the metadata. For instance, for the Edinburgh’s MT tool above it is `/api/elg/v1`.
 - An ID that will be used to call it.
 - Which “try out” UI will be used for testing it and visualizing the returned results.
- **LT Service is tested:** On the LT landing page, there is a “Try out” tab and a “Code samples” tab; both can be used to test the service with some input; see [Test an LT service](#) section. The reviewer can help you identify integration issues and resolve them. This process is continued until the LT service is correctly integrated to the platform. The procedure may require access to the k8s cluster for the reviewer (e.g., to check containers start-up/failures, logs, etc.).
 - **LT Service is published:** When the LT service works as expected, the reviewer will approve it; the metadata record is then published and visible to all ELG users through the catalogue.

6.5 Frequently Asked Questions

Question: What is a k8s namespace and when should an LT Provider ask for one?

Answer: A k8s namespace is a virtual sub-cluster, which can be used to restrict access to the respective containers that run within it. You should ask for a dedicated namespace (in ELG k8s cluster) when you need to ensure isolation and security; i.e, limit access to your container, logs etc.

Question: The image that I have created is not publicly available. Is it possible to register it to the ELG platform?

Answer: Yes, it can be registered. A k8s secret containing the required credentials will be created for the namespace in which your image is going to be deployed. k8s will then be able to pull the image and deploy it.

Question: Are there any requirements for `executionLocation`? For example, an IE tool has to expose a specific path or use a specific port?

Answer: No, you can use any valid port or path. This holds for any kind of LT tool (IE, MT, ASR, etc.). The internal container port will be mapped (via port mapping) to port 80. Remember that the endpoint of the LT service follows this pattern: `http://{k8s service name for the registered LT tool}.{k8s namespace for the registered LT tool}.svc.cluster.local{the path where the REST service is running at}`, which assumes that the service is exposed to port 80.

Question: I have *n* different versions of the same IE LT tool; e.g., one version per language. How should I register them to the platform? I have to create one Docker image with all the different versions or one image per version?

Answer: Both are possible. In both cases you will have to provide a separate metadata record for each LT tool. However, in the case where the tools are packaged together, **all** metadata records must point to the same image location (`dockerDownloadLocation`) and each of them has to listen in a different HTTP endpoint (`executionLocation`) but on the same port (for simplicity). E.g., `“http://localhost:8080/NamedEntityRecognitionEN”`, `“http://localhost:8080/NamedEntityRecognitionDE”`.

Question: Should the Docker image that I will provide have a specific tag?



[Upload](#)

[Technologies](#)

[Resources](#)

[Community](#)

[Events](#)

[About E](#)


[← Back to catalogue](#)



ILSP Named Entity Recognizer for English

ILSP-NER-English

Version: v1.0.0

 ToolService

[Overview](#)

[Download/Run](#)

[Test/Try out](#)

[Code samples](#)

Type text to annotate

I moved to Berlin

SUBMIT

Answer: The images that are stored in GitLab or DockerHub are not immutable, even when they have been assigned a specific/custom tag; thus, it is possible that they are overwritten (by their creators). ELG (currently) does not have a private Docker registry that caches images. Therefore, when ELG will try (at some point) to spawn a new instance of an LT service, it might download (pull) and use an image that is not (any more) ELG compatible, because it has been overwritten (e.g. by accident). So, yes, it is recommended (but not enforced) to put a custom tag (dedicated for ELG) to the image that you will register, since it is usually more common to override the `:latest` one.

Question: How many resources will be allocated for my LT container in the k8s cluster?

Answer: By default, 512MB of RAM and half a CPU core. If your LT service requires more resources you have to specify it by using the `additionalHWRequirements` metadata element (see the MT example above) or by communicating with the ELG administrators.

Provide a Language Resource

You can register at the ELG platform data resources, such as corpora (raw and annotated), computational lexica, terminological glossaries, models, computational grammars, etc. For more information on the resource types, see [Overview](#).

Corpora are structured collections of data selected according to specific criteria in order to represent as comprehensively as possible a research question. The most common cases are:

- text corpora: monolingual, bilingual or multilingual collections of texts in a specific domain, such as corpora of news articles, scientific publications, legal documents, medical records, tweets, etc.
- corpora of audio recordings, e.g., of broadcast news, or lists of sentences recorded by individuals from a specific region with a dialect accent, etc.
- collections of videos, such as interviews with politicians, sign language corpora, etc.
- corpora combining all of the above, such as a multimedia corpus of video lectures, with their audio recordings, transcripts, subtitles and their translations.

Under **language descriptions**, we comprise:

- models, including Machine Learning models, statistical models, word embeddings, n-gram models,
- computational grammars of a language, language variety or for a specific domain or phenomenon.

The vast majority of these consist of a text part, but videos and images are also foreseen for cases such as sign language grammars.

Examples of **lexical/conceptual resources** include

- computational lexica, that are used for computational processing, and include morphological, syntactic and semantic information;
- dictionaries in digital format,
- ontologies and controlled vocabularies,
- monolingual and multilingual terminological glossaries,
- word lists, gazetteers of place names, proper names, etc.

They typically consist of a text part, but they may also comprise audio and video files, as in the case of:

- multimedia lexica with sound recordings (e.g., pronunciation of a word) and images (e.g. pictures denoting the sense of a word),
- sign language lexica with videos.

7.1 Technical requirements

All data resources must be provided as `.zip`, `.tar` or `.tar.gz` archives.

7.2 Describe a Language Resource

To register your resource at ELG, you must describe it according to the *ELG metadata schema* (at least minimal version), i.e., you have to provide a *metadata record*, and upload this description to the platform.

Note: For this release, you **MUST** provide an ELG-compliant XML file. Upcoming releases will also include a metadata editor and other functionalities supporting an easy import of metadata records.

You will find the full schema XSD, documentation and templates and examples of metadata records for all resource types [here](#) and some examples of already registered data resources [here](#).

Examples of metadata records and a list of the metadata elements of the minimal version are given in separate sections:

- corpora (datasets): *Describe a corpus (dataset)*
- language descriptions: *Describe a language description (model, grammar)*
- lexical/conceptual resources: *Describe a lexical/conceptual resource (lexicon, glossary, ontology, etc.)*

7.3 Register a language resource to the platform

The following steps should be followed:

- **Provide a metadata record:** Sign into the ELG platform using your credentials and press the “upload” button on the main menu.

Then upload the XML file that contains the metadata. In the current release this is the only way to provide them. Upcoming releases will also include a metadata editor and other functionalities supporting an easy import of metadata records.

The metadata record is validated at import against the metadata schema. Additional rules that check for syntactic and partial semantic integrity are also used. If the file is found invalid, you will see a message with a list of errors; you must correct them and re-upload the file. If it is valid you will be shown a success message; the file will be imported in the database. At this stage, it is visible only to the platform administrators.

- **Resource is checked:** The administrator will assign it to a reviewer; during the review process, the metadata record is visible only to you (LR provider) and the reviewer.

The screenshot shows the European Language Grid Catalogue interface. At the top, the logo 'EUROPEAN LANGUAGE GRID' is displayed with an 'ALPHA' badge. Navigation links include Administration, Upload, Technologies, Resources, Community, Events, and About ELG. A search bar contains the text 'geolocator' and a 'Search' button. Below the search bar, a 'Clear all filters' button is visible. The search results are categorized on the left under 'Language resources & technologies', 'Languages', 'Service functions', and 'Licences'. The main results area shows '1 search results for geolocator'. The result is for a tool named 'geolocator', described as 'Social media geolocation prediction for a given Tweet, or a short text.' It includes keywords: Geolocation, Social Media, Twitter; language: English; and licence: GNU General Public License v3.0 only. A blue upward arrow icon is next to the result, and a grey 'ingested' badge is present.

- **LR is published:** When the LR has been checked, the reviewer will approve it; the metadata record is then published and visible to all ELG users through the catalogue.

Describe a corpus (dataset)

In this section you will find information on how to describe a corpus with the minimal metadata in order to register it into the ELG platform. If you want to find more on the ELG resource types, see *Overview*. You will also find instructions for all data resources (technical requirements, registration instructions to the platform) in *Provide a Language Resource*.

Corpora are structured collections of data selected according to specific criteria in order to represent as comprehensively as possible a research question. The most common cases are:

- text corpora: monolingual, bilingual or multilingual collections of texts in a specific domain, such as corpora of news articles, scientific publications, legal documents, medical records, tweets, etc.
- corpora of audio recordings, e.g., of broadcast news, or lists of sentences recorded by individuals from a specific region with a dialect accent, etc.
- collections of videos, such as interviews with politicians, sign language corpora, etc.
- corpora combining all of the above, such as a multimedia corpus of video lectures, with their audio recordings, transcripts, subtitles and their translations.

8.1 Examples of metadata records for corpora

Bilingual raw corpus: Bilingual Bulgarian-English corpus from the National Revenue Agency (BG) (Processed)

Published at: <https://live.european-language-grid.eu/catalogue/#/resource/service/corpus/734>

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns="http://w3id.org/meta-share/meta-share/" xmlns:datacite=
↪ "http://purl.org/spar/datacite/" xmlns:dc="http://www.w3.org/ns/dcat#" xmlns:ms=
↪ "http://w3id.org/meta-share/meta-share/" xmlns:omtd="http://w3id.org/meta-share/
↪ omtd-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
↪ xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../Schema/ELG-SHARE.
↪ xsd">
<ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.org/meta-
↪ share/meta-share/elg">value automatically assigned - leave as is</
↪ ms:MetadataRecordIdentifier>
```

(continues on next page)

(continued from previous page)

```

<ms:metadataCreationDate>2020-10-03</ms:metadataCreationDate>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Smith</ms:surname>
  <ms:givenName xml:lang="en">John</ms:givenName>
  <ms:email>username@someDomain.com</ms:email>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Smith</ms:surname>
  <ms:givenName xml:lang="en">John</ms:givenName>
  <ms:email>username@someDomain.com</ms:email>
</ms:metadataCreator>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="en">Bilingual Bulgarian-English_
↪corpus from the National Revenue Agency (BG) (Processed)</ms:resourceName>
    <ms:description xml:lang="en">Bilingual Bulgarian-English_
↪corpus of administrative documents on the Refund of Value Added Tax from the_
↪Bulgarian National Revenue Agency.

    Bilingual Bulgarian-English corpus of administrative documents on the Refund_
↪of Value Added Tax from the Bulgarian National Revenue Agency. It was offered as_
↪collection of documents by the Bulgarian National Revenue Agency. Modules of the_
↪ILSP Focused Crawler was used for the normalization, cleaning, (near) de-
↪duplication and identification of parallel documents. The Maligna sentence aligner_
↪was used for extracting segment alignments from crawled parallel documents. As a_
↪post-processing step, alignments were merged into one TMX file. The following_
↪filters were applied: TMX files generated from document pairs which have been_
↪identified by non-aupidh methods were discarded ; TMX files with a zeroToOne_
↪alignments/total_alignments ratio larger than 0.16, were discarded ; Alignments of_
↪non-[1:1] type(s) were discarded. ; Alignments with a TUV (after normalization)_
↪that has less than 1 tokens, were annotated ; Alignments with a 11/12 TUV length_
↪ratio smaller than 0.6 or larger than 1.6, were annotated ; Alignments in which_
↪different digits appear in each TUV were kept and annotated. ; Alignments with_
↪identical TUVs (after normalization) were annotated ; Alignments with only non-
↪letters in at least one of their TUVs were annotated ; Duplicate alignments were_
↪kept and were annotated. The mean value of aligner's scores is 5.714609036504669,_
↪the std value is 1.8063256236105307. The mean value of length (in terms of_
↪characters) ratios is 1.0040012545201242 and the std value is 0.26545877788005745._
↪There are 832 TUs with no annotation, containing 13336 words and 2604 lexical types_
↪in bul and 15010 words and 2031 lexical types in eng. The mean value of aligner's_
↪scores is 6.336834960545485, the std value is 1.53829791384023</ms:description>
    <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-
↪share/meta-share/other">ELRC_471</ms:LRIdentifier>
    <ms:version>2.0</ms:version>
    <ms:additionalInfo>
      <ms:landingPage>https://elrc-share.eu/repository/
↪browse/bilingual-bulgarian-english-corpus-from-the-national-revenue-agency-bg-
↪processed/4ed47824d04a11e7b7d400155d026706dbe4fc9f12424b5ba0a749fd6758072b/</
↪ms:landingPage>
    </ms:additionalInfo>
    <ms:additionalInfo>
      <ms:email>contact@someDomain.com</ms:email>

```

(continues on next page)

(continued from previous page)

```

        </ms:additionalInfo>
        <ms:contact>
            <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Rusinova</ms:surname>
                <ms:givenName xml:lang="en">Annie</ms:givenName>
                <ms:email>contact@someDomain.com</ms:email>
            </ms:Person>
        </ms:contact>
        <ms:iprHolder>
            <ms:Organization>
                <ms:actorType>Organization</ms:actorType>
                <ms:organizationName xml:lang="en">National Revenue Agency (BG)</ms:organizationName>
                <ms:website>http://www.nap.bg/en/</ms:website>
            </ms:Organization>
        </ms:iprHolder>
        <ms:keyword xml:lang="en">corpus</ms:keyword>
        <ms:domain>
            <ms:categoryLabel xml:lang="en">FINANCE</ms:categoryLabel>
            <ms:DomainIdentifier ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/EUROVOC">24</ms:DomainIdentifier>
        </ms:domain>
        <ms:fundingProject>
            <ms:projectName xml:lang="en">European Language Resource Coordination LOT3</ms:projectName>
            <ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.org/meta-share/meta-share/other">Tools and Resources for CEF Automated Translation - LOT3 (SMART 2015/1091 - 30-CE-0816766/00-92)</ms:ProjectIdentifier>
            <ms:website>http://www.lr-coordination.eu</ms:website>
        </ms:fundingProject>
        <ms:validated>true</ms:validated>
        <ms:validation>
            <ms:validationDetails xml:lang="en">validated</ms:validationDetails>
        </ms:validation>
        <ms:relation>
            <ms:relationType xml:lang="en">isAlignedVersionOf</ms:relationType>
            <ms:relatedLR>
                <ms:resourceName xml:lang="en">Bilingual Bulgarian-English corpus from the National Revenue Agency (BG)</ms:resourceName>
                <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-share/meta-share/other">ELRC_447</ms:LRIdentifier>
            </ms:relatedLR>
        </ms:relation>
        <ms:LRS subclass>
            <ms:Corpus>
                <ms:lrType>Corpus</ms:lrType>
                <ms:corpusSubclass>http://w3id.org/meta-share/meta-share/rawCorpus</ms:corpusSubclass>
                <ms:CorpusMediaPart>
                    <ms:CorpusTextPart>
                        <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
                    </ms:CorpusTextPart>
                </ms:CorpusMediaPart>
            </ms:Corpus>
        </ms:LRS subclass>
    </ms:Corpus>

```

(continues on next page)

(continued from previous page)

```

↪meta-share/meta-share/text</ms:mediaType>
↪w3id.org/meta-share/meta-share/bilingual</ms:lingualityType>
↪w3id.org/meta-share/meta-share/parallel</ms:multilingualityType>
↪ms:languageTag>
↪ms:languageId>
↪ms:scriptId>
↪ms:languageTag>
↪ms:languageId>
↪xml:lang="en">administrativeTexts</ms:categoryLabel>
↪xml:lang="en">official</ms:categoryLabel>
↪org/meta-share/meta-share/mixed</ms:creationMode>
↪xml:lang="en">ELRC-447</ms:resourceName>
↪ms:LRIdentifierScheme="http://w3id.org/meta-share/meta-share/other">ELRC_447</ms:LRIdentifier>
↪"en">See description for creation details</ms:creationDetails>
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
↪eu/repository/download/
↪4ed47824d04a11e7b7d400155d026706dbe4fc9f12424b5ba0a749fd6758072b/</ms:accessLocation>
↪ms:amount>
↪w3id.org/meta-share/meta-share/unit</ms:sizeUnit>
↪org/meta-share/omtd-share/Xml</ms:dataFormat>
<ms:mediaType>http://w3id.org/
<ms:lingualityType>http://
<ms:multilingualityType>http://
<ms:language>
  <ms:languageTag>bg</
  <ms:languageId>bg</
  <ms:scriptId>Cyril</
</ms:language>
<ms:language>
  <ms:languageTag>en</
  <ms:languageId>en</
</ms:language>
<ms:textType>
  <ms:categoryLabel_
</ms:textType>
<ms:TextGenre>
  <ms:categoryLabel_
</ms:TextGenre>
<ms:creationMode>http://w3id.
<ms:hasOriginalSource>
  <ms:resourceName_
  <ms:LRIdentifier_
</ms:hasOriginalSource>
<ms:creationDetails xml:lang=
</ms:CorpusTextPart>
</ms:CorpusMediaPart>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
  <ms:accessLocation>https://elrc-share.
<ms:distributionTextFeature>
  <ms:size>
    <ms:amount>1292</
    <ms:sizeUnit>http://
  </ms:size>
  <ms:dataFormat>http://w3id.

```

(continues on next page)

(continued from previous page)

```

<ms:characterEncoding>http://
↪w3id.org/meta-share/meta-share/UTF-8</ms:characterEncoding>
</ms:distributionTextFeature>
<ms:licenceTerms>
  <ms:licenceTermsName xml:lang=
↪"en">publicDomain</ms:licenceTermsName>
  <ms:licenceTermsURL>https://
↪elrc-share.eu/terms/publicDomain.html</ms:licenceTermsURL>
  <ms:LicenceIdentifier
↪ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">publicDomain
↪</ms:LicenceIdentifier>
</ms:licenceTerms>
<ms:cost>
  <ms:amount>0</ms:amount>
  <ms:currency>http://w3id.org/
↪meta-share/meta-share/euro</ms:currency>
</ms:cost>
</ms:DatasetDistribution>
<ms:personalDataIncluded>false</
↪ms:personalDataIncluded>
<ms:sensitiveDataIncluded>false</
↪ms:sensitiveDataIncluded>
</ms:Corpus>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

Annotated corpus: Greek Textual Entailment corpusPublished at: <https://live.european-language-grid.eu/catalogue/#/resource/service/corpus/649>

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns="http://w3id.org/meta-share/meta-share/" xmlns:datacite=
↪"http://purl.org/spar/datacite/" xmlns:dc="http://www.w3.org/ns/dcat#" xmlns:ms=
↪"http://w3id.org/meta-share/meta-share/" xmlns:omtd="http://w3id.org/meta-share/
↪omtd-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
↪xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../Schema/ELG-SHARE.
↪xsd">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↪org/meta-share/meta-share/elg">value automatically assigned - leave as is</
↪ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-02-02</ms:metadataCreationDate>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>curator@somedomain.com</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>

```

(continues on next page)

(continued from previous page)

```

        <ms:email>curator@somedomain.com</ms:email>
    </ms:metadataCreator>
    <ms:sourceOfMetadataRecord>META-SHARE/ILSP</ms:sourceOfMetadataRecord>
    <ms:DescribedEntity>
        <ms:LanguageResource>
            <ms:entityType>LanguageResource</ms:entityType>
            <ms:resourceName xml:lang="en">Greek Textual Entailment Corpus
↪ </ms:resourceName>
            <ms:resourceShortName xml:lang="en">GTEC</
↪ ms:resourceShortName>
            <ms:description xml:lang="en">GTEC consits of 600 T-H pairs_
↪ manually annotated for entailment (i.e. whether T entails H or not) by human_
↪ annotators. The dataset which is tailored to guide training and evaluation of_
↪ prospect RTE systems, is equally divided in three subsets each one representing the_
↪ output of a specific HLT application: Question Answering (QA), Comparable Documents_
↪ (CD) and Machine Translation (MT), and pertaining to specific subject fields (e.g._
↪ law, politics, travel). T-H examples that correspond to success and failure cases_
↪ of the afore-mentioned applications have been included in the corpus. The_
↪ annotations provided are conformant to the RTE1 and RTE2 challenges.</
↪ ms:description>
                <ms:version>v1.0.0 (automatically assigned)</ms:version>
                <ms:additionalInfo>
                    <ms:email>username@someDomain.com</ms:email>
                </ms:additionalInfo>
                <ms:additionalInfo>
                    <ms:email>username3@someDomain.com</ms:email>
                </ms:additionalInfo>
                <ms:contact>
                    <ms:Person>
                        <ms:actorType>Person</ms:actorType>
                        <ms:surname xml:lang="en">Giouli</ms:surname>
↪ <ms:givenName xml:lang="en">Voula</
↪ ms:givenName>
                            <ms:email>username@someDomain.com</ms:email>
                        </ms:Person>
                    </ms:contact>
                    <ms:contact>
                        <ms:Person>
                            <ms:actorType>Person</ms:actorType>
↪ <ms:surname xml:lang="en">Piperidis</
↪ ms:surname>
                                    <ms:givenName xml:lang="en">Stelios</
↪ ms:givenName>
                                            <ms:email>username3@someDomain.com</ms:email>
                                </ms:Person>
                            </ms:contact>
                        <ms:keyword xml:lang="en">corpus</ms:keyword>
                        <ms:domain>
                            <ms:categoryLabel xml:lang="en">law</ms:categoryLabel>
                        </ms:domain>
                        <ms:domain>
                            <ms:categoryLabel xml:lang="en">politics</
↪ ms:categoryLabel>
                                </ms:domain>
                        <ms:domain>
                            <ms:categoryLabel xml:lang="en">travel</
↪ ms:categoryLabel>

```

(continues on next page)

(continued from previous page)

```

</ms:domain>
<ms:resourceCreator>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="en">Institute_
↪for Language and Speech Processing</ms:organizationName>
    <ms:website>http://www.ilsp.gr</ms:website>
  </ms:Organization>
</ms:resourceCreator>
<ms:intendedApplication>
  <ms:LTCClassRecommended>http://w3id.org/meta-share/
↪omtd-share/AnnotationOfTextualEntailment</ms:LTCClassRecommended>
</ms:intendedApplication>
<ms:actualUse>
  <ms:usedInApplication>
    <ms:LTCClassRecommended>http://w3id.org/meta-
↪share/omtd-share/AnnotationOfTextualEntailment</ms:LTCClassRecommended>
  </ms:usedInApplication>
  <ms:actualUseDetails xml:lang="en">nlpApplications</
↪ms:actualUseDetails>
    </ms:actualUse>
    <ms:isDocumentedBy>
      <ms:title xml:lang="en">Building a Greek corpus of_
↪Textual Entailment</ms:title>
      <ms:DocumentIdentifier ms:DocumentIdentifierScheme=
↪"http://purl.org/spar/datacite/url">http://www.lrec-conf.org/proceedings/lrec2008/
↪pdf/427_paper.pdf</ms:DocumentIdentifier>
    </ms:isDocumentedBy>
    <ms:LRSubclass>
      <ms:Corpus>
        <ms:lrType>Corpus</ms:lrType>
        <ms:corpusSubclass>http://w3id.org/meta-share/
↪meta-share/annotatedCorpus</ms:corpusSubclass>
        <ms:CorpusMediaPart>
          <ms:CorpusTextPart>
            <ms:corpusMediaType>
↪CorpusTextPart</ms:corpusMediaType>
            <ms:mediaType>http://w3id.org/
↪meta-share/meta-share/text</ms:mediaType>
            <ms:lingualityType>http://
↪w3id.org/meta-share/meta-share/monolingual</ms:lingualityType>
            <ms:language>
              <ms:languageTag>el</
↪ms:languageTag>
              <ms:languageId>el</
↪ms:languageId>
            </ms:language>
            <ms:creationMode>http://w3id.
↪org/meta-share/meta-share/mixed</ms:creationMode>
            <ms:originalSourceDescription_
↪xml:lang="en">web news</ms:originalSourceDescription>
            <ms:originalSourceDescription_
↪xml:lang="en">EU texts</ms:originalSourceDescription>
          </ms:CorpusTextPart>
        </ms:CorpusMediaPart>
      <ms:DatasetDistribution>
        <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>

```

(continues on next page)

(continued from previous page)

```

<ms:accessLocation>http://metashare.
↪ilsp.gr:8080/repository/download/
↪26dca2fe63d211e29b2c842b2b6a04d7db87c85bfbe34326bb4c2e88b8c4da85</ms:accessLocation>
<ms:distributionTextFeature>
  <ms:size>
    <ms:amount>600</
↪ms:amount>
    <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/T-HPair</ms:sizeUnit>
  </ms:size>
  <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Xml</ms:dataFormat>
</ms:distributionTextFeature>
<ms:licenceTerms>
  <ms:licenceTermsName xml:lang=
↪"en">CC-BY-4.0</ms:licenceTermsName>
  <ms:licenceTermsURL>https://
↪spdx.org/licenses/CC-BY-4.0.html</ms:licenceTermsURL>
</ms:licenceTerms>
  <ms:attributionText xml:lang="en">
↪Greek Textual Entailment Corpus by Athena R.C./ILSP used under CC-BY licence</
↪ms:attributionText>
</ms:DatasetDistribution>
<ms:personalDataIncluded>>false</
↪ms:personalDataIncluded>
<ms:sensitiveDataIncluded>>false</
↪ms:sensitiveDataIncluded>
<ms:annotation>
  <ms:annotationType>http://w3id.org/
↪meta-share/omtd-share/Lemma</ms:annotationType>
  <ms:annotationStandoff>>false</
↪ms:annotationStandoff>
  <ms:annotationMode>http://w3id.org/
↪meta-share/meta-share/mixed</ms:annotationMode>
  <ms:annotationModeDetails xml:lang="en
↪">automatic annotation followed with manual disambiguation</
↪ms:annotationModeDetails>
  <ms:isAnnotatedBy>
    <ms:resourceName xml:lang="en
↪">ILSP-Lemmatizer</ms:resourceName>
  </ms:isAnnotatedBy>
</ms:annotation>
<ms:annotation>
  <ms:annotationType>http://w3id.org/
↪meta-share/omtd-share/PartOfSpeech</ms:annotationType>
  <ms:annotationStandoff>>false</
↪ms:annotationStandoff>
  <ms:tagset>
    <ms:resourceName xml:lang="en
↪">ILSP/PAROLE tagset</ms:resourceName>
  </ms:tagset>
  <ms:annotationMode>http://w3id.org/
↪meta-share/meta-share/mixed</ms:annotationMode>
  <ms:annotationModeDetails xml:lang="en
↪">automatic annotation followed with manual disambiguation</
↪ms:annotationModeDetails>
  <ms:isAnnotatedBy>

```

(continues on next page)

(continued from previous page)

```

<ms:resourceName xml:lang="en
↪">ILSP FBT POS tagger</ms:resourceName>
                                </ms:isAnnotatedBy>
                                </ms:annotation>
                                <ms:annotation>
                                <ms:annotationType>http://w3id.org/
↪meta-share/omtd-share/SyntacticAnnotationType</ms:annotationType>
                                <ms:annotationStandoff>>false</
↪ms:annotationStandoff>
                                <ms:annotationMode>http://w3id.org/
↪meta-share/meta-share/mixed</ms:annotationMode>
                                </ms:annotation>
                                <ms:annotation>
                                <ms:annotationType>http://w3id.org/
↪meta-share/omtd-share/SyntacticAnnotationType</ms:annotationType>
                                <ms:annotationStandoff>>false</
↪ms:annotationStandoff>
                                <ms:annotationMode>http://w3id.org/
↪meta-share/meta-share/mixed</ms:annotationMode>
                                <ms:annotationModeDetails xml:lang="en
↪">Automatic annotation followed by manual correction</ms:annotationModeDetails>
                                </ms:annotation>
                                <ms:annotation>
                                <ms:annotationType>http://w3id.org/
↪meta-share/omtd-share/SemanticAnnotationType</ms:annotationType>
                                <ms:annotationStandoff>>false</
↪ms:annotationStandoff>
                                <ms:annotationMode>http://w3id.org/
↪meta-share/meta-share/manual</ms:annotationMode>
                                </ms:annotation>
                                </ms:Corpus>
                                </ms:LRSubclass>
                                </ms:LanguageResource>
                                </ms:DescribedEntity>
</ms:MetadataRecord>

```

8.2 Minimal version metadata for corpora

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources** including data language resources are presented in section *Minimal version - List of elements common to all LRTs*. **In addition**, the metadata elements that are required or recommended for corpora are described below.

For a quick guide to the ELG template, see *Template - Explanations*.

8.2.1 Corpus

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpora

Data type component

Optionality Mandatory

Explanation & Instructions

Wraps together the set of elements that is specific to corpora

Example

```
<ms:LRSubclass>
  <ms:Corpus>
    <ms:lrType>Corpus</ms:lrType>
  </ms:Corpus>
</ms:LRSubclass>
```

8.2.2 corpusSubclass

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpora.corpusSubclass

Data type CV ([corpusSubclass](#))

Optionality Mandatory

Explanation & Instructions

Introduces a classification of corpora into types (used for descriptive reasons)

Use one of the values for raw corpora, annotated corpora (mixed raw with annotations), annotations (only annotations without the original corpus)

Example

```
<ms:corpusSubclass>http://w3id.org/meta-share/meta-share/rawCorpus</ms:corpusSubclass>

<ms:corpusSubclass>http://w3id.org/meta-share/meta-share/annotatedCorpus</
<ms:corpusSubclass>
```

8.2.3 CorpusTextPart

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpora.CorporaMediaPart.CorporaTextPart

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

The part of a corpus (or a whole corpus) that consists of textual segments (e.g., a corpus of publications, or transcriptions of an oral corpus, or subtitles, etc.)

You can repeat the group of elements for multiple textual parts.

The mandatory or recommended elements for the text part are:

- **mediaType** (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value 'text'.
- **lingualityType** (Mandatory): Indicates whether the resource includes one, two or more languages.

- `multilingualityType` (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If `lingualityType` = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and `multilingualSingleText` (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language).
- `language` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `languageVariety` (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.
- `modalityType` (Recommended if applicable): Specifies the type of the modality represented in the resource. For instance, you can use 'spoken language' to describe transcribed speech corpora.
- `TextGenre` (Recommended): A category of text characterized by a particular style, form, or content according to a specific classification scheme. See [TextGenre](#).

Example

```

<ms:CorpusTextPart>
  <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>es</ms:languageTag>
    <ms:languageId>es</ms:languageId>
  </ms:language>
</ms:CorpusTextPart>

<ms:CorpusTextPart>
  <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/bilingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>es</ms:languageTag>
    <ms:languageId>es</ms:languageId>
  </ms:language>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>
  </ms:language>
  <ms:multilingualityType>http://w3id.org/meta-share/meta-share/parallel</
↪ms:multilingualityType>
  <ms:TextGenre>
    <ms:CategoryLabel>administrative texts</ms:CategoryLabel>
  </ms:TextGenre>
</ms:CorpusTextPart>

<ms:CorpusTextPart>
  <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>

```

(continues on next page)

(continued from previous page)

```

</ms:language>
<ms:modalityType>http://w3id.org/meta-share/meta-share/spokenLanguage</
↪ms:modalityType>
</ms:CorpusTextPart>

```

8.2.4 CorpusAudioPart

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusAudioPart

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

The part of a corpus (or whole corpus) that consists of audio segments

You can repeat the group of elements for multiple audio parts.

The mandatory or recommended elements for the audio part are:

- *mediaType* (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value ‘audio’
- *lingualityType* (Mandatory): Indicates whether the resource includes one, two or more languages
- *multilingualityType* (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If *lingualityType* = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and multilingualSingleText (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language)
- *language* (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#)
- *languageVariety* (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.
- *modalityType* (Recommended if applicable): Specifies the type of the modality represented in the resource. For instance, you can use ‘spoken language’ to describe transcribed speech corpora.
- *AudioGenre* (Recommended if applicable): A category of audio characterized by a particular style, form, or content according to a specific classification scheme. See [AudioGenre](#)
- *SpeechGenre* (Recommended if applicable): A category for the conventionalized discourse of the speech part of a language resource, based on extra-linguistic and internal linguistic criteria. See [SpeechGenre](#)

Example

```

<ms:CorpusAudioPart>
  <ms:corpusMediaType>CorpusAudioPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/audio</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>

```

(continues on next page)

(continued from previous page)

```

        </ms:language>
        <ms:AudioGenre>
            <ms:CategoryLabel>conference noises</ms:CategoryLabel>
        </ms:AudioGenre>
    </ms:CorpusAudioPart>

    <ms:CorpusAudioPart>
        <ms:corpusMediaType>CorpusAudioPart</ms:corpusMediaType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/audio</ms:mediaType>
        <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
    <ms:lingualityType>
        <ms:language>
            <ms:languageTag>en</ms:languageTag>
            <ms:languageId>en</ms:languageId>
        </ms:language>
        <ms:modalityType>http://w3id.org/meta-share/meta-share/spokenLanguage</
    <ms:modalityType>
        <ms:SpeechGenre>
            <ms:CategoryLabel>monologue</ms:CategoryLabel>
        </ms:SpeechGenre>
    </ms:CorpusAudioPart>

```

8.2.5 CorpusVideoPart

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusVideoPart

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

The part of a corpus (or a whole corpus) that consists of video segments (e.g., a corpus of video lectures, a part of a corpus with news, a sign language corpus, etc.)

You can repeat the group of elements for multiple video parts.

The mandatory or recommended elements for the video part are:

- `mediaType` (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value ‘video’.
- `lingualityType` (Mandatory): Indicates whether the resource includes one, two or more languages.
- `multilingualityType` (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If `lingualityType` = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and multilingualSingleText (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language).
- `language` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `languageVariety` (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.

- `modalityType` (Recommended if applicable): Specifies the type of the modality represented in the resource. For instance, you can use ‘spoken language’ to describe transcribed speech corpora.
- `VideoGenre` (Recommended): A classification of video parts based on extra-linguistic and internal linguistic criteria and reflected on the video style, form or content. See [VideoGenre](#)
- `typeOfVideoContent` (Mandatory): Main type of object or people represented in the video.

Example

```
<ms:CorpusVideoPart>
  <ms:corpusMediaType>CorpusVideoPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/video</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>
  </ms:language>
  <ms:modalityType>http://w3id.org/meta-share/meta-share/bodyGesture</
↪ms:modalityType>
  <ms:modalityType>http://w3id.org/meta-share/meta-share/facialExpression</
↪ms:modalityType>
  <ms:modalityType>http://w3id.org/meta-share/meta-share/spokenLanguage</
↪ms:modalityType>
  <ms:typeOfVideoContent>people eating at a restaurant</ms:typeOfVideoContent>
</ms:CorpusVideoPart>

<ms:CorpusVideoPart>
  <ms:corpusMediaType>CorpusVideoPart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/video</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
  <ms:language>
    <ms:languageTag>fr</ms:languageTag>
    <ms:languageId>fr</ms:languageId>
  </ms:language>
  <ms:VideoGenre>
    <ms:CategoryLabel>documentary</ms:CategoryLabel>
  </ms:VideoGenre>
  <ms:typeOfVideoContent>birds, wild animals, plants</ms:typeOfVideoContent>
</ms:CorpusVideoPart>
```

8.2.6 CorpusImagePart

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusImagePart`

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

The part of a corpus (or whole corpus) that consists of images (e.g., g a corpus of photographs and their captions)

You can repeat the group of elements for multiple video parts.

The mandatory or recommended elements for the image part are:

- `mediaType` (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value 'image'.
- `lingualityType` (Mandatory): Indicates whether the resource includes one, two or more languages.
- `multilingualityType` (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If `lingualityType` = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and `multilingualSingleText` (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language).
- `language` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `languageVariety` (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.
- `modalityType` (Recommended if applicable): Specifies the type of the modality represented in the resource.
- `ImageGenre` (Recommended): A category of images characterized by a particular style, form, or content according to a specific classification scheme. See [ImageGenre](#).
- `typeOfImageContent` (Mandatory): Main type of object or people represented in the image.

Example

```
<ms:CorpusImagePart>
  <ms:corpusMediaType>CorpusImagePart</ms:corpusMediaType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/image</ms:mediaType>
  <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
  <ms:lingualityType>
    <ms:language>
      <ms:languageTag>el</ms:languageTag>
      <ms:languageId>el</ms:languageId>
    </ms:language>
    <ms:ImageGenre>
      <ms:CategoryLabel>comics</ms:CategoryLabel>
    </ms:ImageGenre>
    <ms:typeOfImageContent>human figures</ms:typeOfImageContent>
</ms:CorpusImagePart>
```

8.2.7 TextGenre

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpora.CorporaMediaPart.CorporaTextPart.TextGenre`

Data type component

Optionality Recommended

Explanation & Instructions

A category of text characterized by a particular style, form, or content according to a specific classification scheme

You can add only a free text value at the `CategoryLabel` element; if you have used a value from an established controlled vocabulary, you can use the `TextGenreIdentifier` and the attribute `TextGenreClassificationScheme`.

Example

```
<ms:TextGenre>
  <ms:CategoryLabel>movie subtitles</ms:CategoryLabel>
</ms:TextGenre>

<ms:TextGenre>
  <ms:CategoryLabel>news articles</ms:CategoryLabel>
</ms:TextGenre>
```

8.2.8 AudioGenre

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusAudioPart

Data type component

Optionality Recommended if applicable

Explanation & Instructions

A category of audio characterized by a particular style, form, or content according to a specific classification scheme

You can add only a free text value at the `CategoryLabel` element; if you have used a value from an established controlled vocabulary, you can use the `AudioGenreIdentifier` and the attribute `AudioGenreClassificationScheme` to provide further details.

Example

```
<ms:AudioGenre>
  <ms:CategoryLabel>conference noises</ms:CategoryLabel>
</ms:AudioGenre>
```

8.2.9 SpeechGenre

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusAudioPart.SpeechGenre

Data type component

Optionality Recommended if applicable

Explanation & Instructions

A category for the conventionalized discourse of the speech part of a language resource, based on extra-linguistic and internal linguistic criteria

You can add only a free text value at the `CategoryLabel` element; if you have used a value from an established controlled vocabulary, you can use the `SpeechGenreIdentifier` and the attribute `SpeechGenreClassificationScheme` to provide further details.

Example

```

<ms:SpeechGenre>
  <ms:CategoryLabel>broadcast news</ms:CategoryLabel>
</ms:SpeechGenre>

<ms:SpeechGenre>
  <ms:CategoryLabel>monologue</ms:CategoryLabel>
</ms:SpeechGenre>

```

8.2.10 VideoGenre

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusVideoPart.VideoGenre

Data type string (+ id + scheme)

Optionality Recommended if applicable

Explanation & Instructions

A classification of video parts based on extra-linguistic and internal linguistic criteria and reflected on the video style, form or content

You can add only a free text value at the `CategoryLabel` element; if you have used a value from an established controlled vocabulary, you can use the `VideoGenreIdentifier` and the attribute `VideoClassificationScheme`

Example

```

<ms:videoGenre>
  <ms:CategoryLabel>documentaries</ms:CategoryLabel>
</ms:videoGenre>

<ms:videoGenre>
  <ms:CategoryLabel>video lectures</ms:CategoryLabel>
</ms:videoGenre>

```

8.2.11 ImageGenre

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.CorporusImagePart.ImageGenre

Data type component

Optionality Recommended

Explanation & Instructions

A category of images characterized by a particular style, form, or content according to a specific classification scheme

You can add only a free text value at the `CategoryLabel` element; if you have used a value from an established controlled vocabulary, you can use the `ImageGenreIdentifier` and the attribute `ImageClassificationScheme` to provide further details.

Example

```
<ms:imageGenre>
  <ms:CategoryLabel>human faces</ms:CategoryLabel>
</ms:imageGenre>

<ms:imageGenre>
  <ms:CategoryLabel>landscape</ms:CategoryLabel>
</ms:imageGenre>
```

8.2.12 DatasetDistribution

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.DatasetDistribution`

Data type component

Optionality Mandatory

Explanation & Instructions

Any form with which a dataset is distributed, such as a downloadable form in a specific format (e.g., spreadsheet, plain text, etc.) or an API with which it can be accessed

You can repeat the element for multiple distributions.

The list of mandatory and recommended elements are:

- `DatasetDistributionForm` (Mandatory): The form (medium/channel) used for distributing a language resource consisting of data (e.g., a corpus, a lexicon, etc.). The typical values are ‘downloadable’, ‘accessibleThroughInterface’, ‘accessibleThroughQuery’ (see more at [DatasetDistributionForm](#)).
- `downloadLocation` (Mandatory if applicable): A URL where the language resource (mainly data but also downloadable software programmes or forms) can be downloaded from. Use this element if the value of `DatasetDistributionForm` is ‘downloadable’ and only for direct download links (i.e., from which the dataset is downloaded without the need of further actions such as clicks on a page).
- `accessLocation` (Mandatory if applicable): A URL where the resource can be accessed from; it can be used for landing pages or for cases where the resource is accessible via an interface, i.e. cases where the resource itself is not provided with a direct link for downloading. Use if the value of `DatasetDistributionForm` is ‘accessibleThroughInterface’ or ‘accessibleThroughQuery’ but also for links used for downloading corpora which are mentioned on a landing page or require some kind of action on the part of the user.
- `samplesLocation` (Recommended): Links a resource to a url (or url’s) with samples of a data resource or of the input of output resource of a tool/service.
- `licenceTerms` (Mandatory): See [licenceTerms](#)
- `cost` (Mandatory if applicable): Introduces the cost for accessing a resource, formally described as a set of amount and currency unit. Please use only for resources available at a cost and not for free resources.

Depending on the parts of the corpus, you must also use one or more of the following:

- `distributionTextFeature`: See [distributionTextFeature](#)
- `distributionAudioFeature`: See [distributionAudioFeature](#)
- `distributionVideoFeature`: See [distributionVideoFeature](#)
- `distributionImageFeature`: See [distributionImageFeature](#)

Example


```

<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/downloadable
↪ </ms:DatasetDistributionForm>
  <ms:accessLocation>https://www.someAccessURL.com</ms:accessLocation>
  <ms:samplesLocation>https://www.URLwithsamples.com</ms:samplesLocation>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>17601</ms:amount>
      <ms:sizeUnit>http://w3id.org/meta-share/meta-share/unit</
↪ ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
↪ ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ ms:characterEncoding>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang="en">openUnder-PSI</ms:licenceTermsName>
    <ms:licenceTermsURL>https://elrc-share.eu/terms/openUnderPSI.html</
↪ ms:licenceTermsURL>
  </ms:licenceTerms>
</ms:DatasetDistribution>

<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪ accessibleThroughInterface</ms:DatasetDistributionForm>
  <ms:accessLocation>https://www.someAccessURL.com</ms:accessLocation>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>100</ms:amount>
      <ms:sizeUnit>http://w3id.org/meta-share/meta-share/text1</
↪ ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf</
↪ ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ ms:characterEncoding>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang="en">some commercial licence</
↪ ms:licenceTermsName>
    <ms:licenceTermsURL>https://elrc-share.eu/terms/someCommercialLicence.
↪ html</ms:licenceTermsURL>
  </ms:licenceTerms>
  <ms:cost>
    <ms:amount>10000</ms:amount>
    <ms:currency>http://w3id.org/meta-share/meta-share/euro</ms:currency>
  </ms:cost>
</ms:DatasetDistribution>

```

8.2.13 distributionTextFeature

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.
DatasetDistribution.distributionTextFeature

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Links to a feature that can be used for describing distinct distributable forms of text resources/parts

The following are mandatory or recommended:

- `size` (Mandatory): The size of the text part, expressed as a combination of `amount` and `sizeUnit` (with a value from a CV for `sizeUnit`).
- `dataFormat` (Mandatory): Indicates the format(s) of a data resource; it takes a value from a CV (`dataFormat`); the `dataFormat` includes the IANA mimetype and pointers to additional documentation for specialized formats (e.g., GATE XML, CONLL formats, etc.).
- `characterEncoding` (Recommended): Specifies the character encoding used for a language resource data distribution.

Example

```
<ms:distributionTextFeature>
  <ms:size>
    <ms:amount>9139</ms:amount>
    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/sentence</
↪ms:sizeUnit>
  </ms:size>
  <ms:size>
    <ms:amount>40</ms:amount>
    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</ms:sizeUnit>
  </ms:size>
  <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</ms:dataFormat>
  <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
</ms:distributionTextFeature>
```

8.2.14 distributionAudioFeature

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.DatasetDistribution.distributionAudioFeature`

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Links to a feature that can be used for describing distinct distributable forms of audio resources/parts

The following are mandatory or recommended:

- `size` (Mandatory): The size of the audio part, expressed as a combination of `amount` and `sizeUnit` (with a value from a CV for `sizeUnit`).
- `durationOfAudio` (Recommended): Specifies the duration of the audio recording including silences, music, pauses, etc., expressed as a combination of `amount` and `durationUnit` (with a value from the CV for `durationUnit`).

- `durationOfEffectiveSpeech` (Recommended): Specifies the duration of effective speech of the audio (part of a) resource, expressed as a combination of `amount` and `durationUnit` (with a value from the CV for `durationUnit`).
- `audioFormat` (Mandatory): Indicates the format(s) of the audio (part of a) data resource, expressed as a value of `dataFormat` (with a value from a CV) and `compressed`.

Example

```
<ms:distributionAudioFeature>
  <ms:size>
    <ms:amount>10</ms:amount>
    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</ms:sizeUnit>
  </ms:size>
  <ms:durationOfAudio>
    <ms:amount>3</ms:amount>
    <ms:durationUnit>http://w3id.org/meta-share/meta-share/hour</
↪ms:durationUnit>
  </ms:durationOfAudio>
  <ms:audioFormat>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/wav</
↪ms:dataFormat>
    <ms:compressed>true</ms:compressed>
  </ms:audioFormat>
</ms:distributionAudioFeature>
```

8.2.15 distributionVideoFeature

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.DatasetDistribution.distributionVideoFeature`

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Links to a feature that can be used for describing distinct distributable forms of video resources/parts

The following are mandatory or recommended:

- `size` (Mandatory): The size of the video part, expressed as a combination of `amount` and `sizeUnit` (with a value from a CV for `sizeUnit`).
- `durationOfVideo` (Recommended): Specifies the duration of the video recording, expressed as a combination of `amount` and `durationUnit` (with a value from the CV for `durationUnit`).
- `videoFormat` (Mandatory): Indicates the format(s) of the video (part of a) data resource, expressed as a value of `dataFormat` (with a value from a CV) and `compressed`.

Example

```
<ms:distributionVideoFeature>
  <ms:size>
    <ms:amount>9139</ms:amount>
    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/screen</
↪ms:sizeUnit>
  </ms:size>
```

(continues on next page)

(continued from previous page)

```

    <ms:size>
      <ms:amount>40</ms:amount>
      <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</ms:sizeUnit>
    </ms:size>
    <ms:durationOfVideo>
      <ms:amount>40</ms:amount>
      <ms:durationUnit>http://w3id.org/meta-share/meta-share/hour</
↪ms:durationUnit>
    </ms:durationOfVideo>
    <ms:videoFormat>
      <ms:dataFormat>http://w3id.org/meta-share/omtd-share/wav</
↪ms:dataFormat>
      <ms:compressed>true</ms:compressed>
    </ms:videoFormat>

```

8.2.16 distributionImageFeature

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.
DatasetDistribution.distributionImageFeature

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Links to a feature that can be used for describing distinct distributable forms of image resources/parts

The following are mandatory or recommended:

- `size` (Mandatory): The size of the image part, expressed as a combination of `amount` and `sizeUnit` (with a value from a CV for `sizeUnit`).
- `imageFormat` (Mandatory): Indicates the format(s) of the image (part of a) data resource, expressed as a value of `dataFormat` (with a value from a CV) and `compressed`.

Example

```

<ms:distributionImageFeature>
  <ms:size>
    <ms:amount>100</ms:amount>
    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</ms:sizeUnit>
  </ms:size>
  <ms:imageFormat>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf` </
↪ms:dataFormat>
    <ms:compressed>true</ms:compressed>
  </ms:imageFormat>

```

8.2.17 personalDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.
personalDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains personal data (mainly in the sense falling under the GDPR)

If the resource contains personal data, you can use the (optional) `personalDataDetails` to provide more information

Example

```
<ms:personalDataIncluded>true</ms:personalDataIncluded>
<ms:personalDataDetails>The corpus contains data on the place of living and place of
↪birth of participants</ms:personalDataDetails>
```

8.2.18 sensitiveDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.sensitiveDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains sensitive data (e.g., medical/health-related, etc.) and thus requires special handling

If the resource contains sensitive data, you can use the (optional) `sensitiveDataDetails` to provide more information.

Example

```
<ms:sensitiveDataIncluded>true</ms:sensitiveDataIncluded>
<ms:sensitiveDataDetails>The corpus contains medical data for persons with
↪disabilities</ms:sensitiveDataDetails>
```

8.2.19 anonymized

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.anonymized

Data type boolean

Optionality Mandatory if applicable

Explanation & Instructions

Indicates whether the language resource has been anonymized

The element is mandatory if either `personalDataIncluded` or `sensitiveDataIncluded` have ‘true’ as value; `anonymizationDetails` must also be filled in with information on the anonymization method, etc.

Example

```
<ms:anonymized>true</ms:anonymized>
<ms:anonymizationDetails>pseudonymization performed manually</ms:anonymizationDetails>
```

8.2.20 annotation

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.
annotation

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

Links a corpus to its annotated part(s)

You must use it for annotated corpora and annotations. You can repeat it for corpora that have separate files for each annotation type, or if you want to give information such as the use of different annotation tools for each annotation level.

Enter at least the annotation type(s); if you want, you can give a more detailed description of the annotated parts - see the [annotation](#) component of the full schema.

Example

```
<ms:annotation>
  <ms:annotationType>http://w3id.org/meta-share/omtd-share/Lemma</
↪ms:annotationType>
  <ms:annotationStandoff>>false</ms:annotationStandoff>
  <ms:annotationMode>http://w3id.org/meta-share/meta-share/mixed</
↪ms:annotationMode>
  <ms:isAnnotatedBy>
    <ms:resourceName xml:lang="en">Lemmatizer</ms:resourceName>
  </ms:isAnnotatedBy>
</ms:annotation>

<ms:annotation>
  <ms:annotationType>http://w3id.org/meta-share/omtd-share/PartOfSpeech</
↪ms:annotationType>
  <ms:annotationStandoff>>false</ms:annotationStandoff>
  <ms:tagset>
    <ms:resourceName xml:lang="en">Universal Dependencies</
↪ms:resourceName>
  </ms:tagset>
  <ms:isAnnotatedBy>
    <ms:resourceName xml:lang="en">PoS tagger</ms:resourceName>
  </ms:isAnnotatedBy>
</ms:annotation>

<ms:annotation>
  <ms:annotationType>http://w3id.org/meta-share/omtd-share/
↪SyntacticAnnotationType</ms:annotationType>
</ms:annotation>
```

Describe a language description (model, grammar)

In this section you will find information on how to describe a language description (model, grammar) with the minimal metadata in order to register it in the ELG platform. If you want to find more on the ELG resource types, see [Overview](#). You will also find instructions for all data resources (technical requirements, registration instructions to the platform) in [Provide a Language Resource](#).

Under **language descriptions**, we comprise:

- models, including Machine Learning models, statistical models, word embeddings, n-gram models,
- computational grammars of a language, language variety or for a specific domain or phenomenon.

The vast majority of these consist of a text part, but videos and images are also foreseen for cases such as sign language grammars.

9.1 Examples of metadata records for language descriptions

Monolingual computational grammar for a specific domain: Tourism Italian grammar Published at: <https://live.european-language-grid.eu/catalogue/#/resource/service/ld/901>

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns="http://w3id.org/meta-share/meta-share/" xmlns:datacite=
→ "http://purl.org/spar/datacite/" xmlns:dc="http://www.w3.org/ns/dcat#" xmlns:ms=
→ "http://w3id.org/meta-share/meta-share/" xmlns:omtd="http://w3id.org/meta-share/
→ omtd-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
→ xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../Schema/ELG-SHARE.
→ xsd">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
→ org/meta-share/meta-share/elg">value automatically assigned - leave as is</
→ ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-10-03</ms:metadataCreationDate>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
```

(continues on next page)

(continued from previous page)

```

        <ms:givenName xml:lang="en">John</ms:givenName>
        <ms:email>username@someDomain.com</ms:email>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
    <ms:metadataCreator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Smith</ms:surname>
        <ms:givenName xml:lang="en">John</ms:givenName>
        <ms:email>username@someDomain.com</ms:email>
    </ms:metadataCreator>
    <ms:DescribedEntity>
        <ms:LanguageResource>
            <ms:entityType>LanguageResource</ms:entityType>
            <ms:resourceName xml:lang="en">Tourism Italian grammar</
↪ms:resourceName>
            <ms:resourceShortName xml:lang="en">Tour.ita.grm</
↪ms:resourceShortName>
            <ms:description xml:lang="en">Tourism Italian abnf grammar,
↪manually created. Created within the Portdial project</ms:description>
            <ms:version>v1.0.0 (automatically assigned)</ms:version>
            <ms:additionalInfo>
                <ms:landingPage>https://sites.google.com/site/
↪portdial2</ms:landingPage>
            </ms:additionalInfo>
            <ms:additionalInfo>
                <ms:email>contact@someDomain.com</ms:email>
            </ms:additionalInfo>
            <ms:contact>
                <ms:Person>
                    <ms:actorType>Person</ms:actorType>
                    <ms:surname xml:lang="en">Potamianos</
↪ms:surname>
                    <ms:givenName xml:lang="en">Alex</
↪ms:givenName>
                    <ms:email>contact@someDomain.com</ms:email>
                </ms:Person>
            </ms:contact>
            <ms:keyword xml:lang="en">languageDescription</ms:keyword>
            <ms:fundingProject>
                <ms:projectName xml:lang="en">Portdial</
↪ms:projectName>
            </ms:fundingProject>
            <ms:LRSubclass>
                <ms:LanguageDescription>
                    <ms:lrType>LanguageDescription</ms:lrType>
                    <ms:LanguageDescriptionSubclass>
                        <ms:Grammar>
                            <ms:ldSubclassType>Grammar</
↪ms:ldSubclassType>
                            <ms:encodingLevel>http://w3id.
↪org/meta-share/meta-share/morphology</ms:encodingLevel>
                        </ms:Grammar>
                    </ms:LanguageDescriptionSubclass>
                    <ms:LanguageDescriptionMediaPart>
                        <ms:LanguageDescriptionTextPart>
                            <ms:ldMediaType>
↪LanguageDescriptionTextPart</ms:ldMediaType>

```

(continues on next page)

(continued from previous page)

```

↪meta-share/meta-share/text</ms:mediaType>
<ms:mediaType>http://w3id.org/
↪w3id.org/meta-share/meta-share/monolingual</ms:lingualityType>
<ms:lingualityType>http://
<ms:language>
  ↪ms:languageTag>
  <ms:languageTag>it</
  ↪ms:languageId>
  <ms:languageId>it</
</ms:language>
<ms:metalanguage>
  ↪ms:languageTag>
  <ms:languageTag>und</
  ↪ms:languageId>
  <ms:languageId>und</
</ms:metalanguage>
</ms:LanguageDescriptionTextPart>
</ms:LanguageDescriptionMediaPart>
<ms:DatasetDistribution>
  ↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:DatasetDistributionForm>http://
  ↪ms:accessLocation>
  <ms:accessLocation>http://accessURL</
  ↪ms:licenceTerms>
  <ms:licenceTermsName xml:lang=
  ↪"en">CC-BY-SA-4.0</ms:licenceTermsName>
  <ms:licenceTermsURL>https://
  ↪spx.org/licenses/CC-BY-SA-4.0.html</ms:licenceTermsURL>
  </ms:licenceTerms>
</ms:DatasetDistribution>
<ms:personalDataIncluded>false</
  ↪ms:personalDataIncluded>
  <ms:sensitiveDataIncluded>false</
  ↪ms:sensitiveDataIncluded>
</ms:LanguageDescription>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

N-gram model: PANACEA Environment Corpus n-grams EL (Greek) Published at: <https://live.european-language-grid.eu/catalogue/#/resource/service/ld/900>

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns="http://w3id.org/meta-share/meta-share/" xmlns:datacite=
  ↪"http://purl.org/spar/datacite/" xmlns:dc="http://www.w3.org/ns/dc#" xmlns:ms=
  ↪"http://w3id.org/meta-share/meta-share/" xmlns:omtd="http://w3id.org/meta-share/
  ↪omtd-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  ↪xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../Schema/ELG-SHARE.
  ↪xsd">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
  ↪org/meta-share/meta-share/elg">value automatically assigned - leave as is</
  ↪ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-10-03</ms:metadataCreationDate>
  <ms:metadataCurator>

```

(continues on next page)

(continued from previous page)

```

    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>username@someDomain.com</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>username@someDomain.com</ms:email>
  </ms:metadataCreator>
  <ms:DescribedEntity>
    <ms:LanguageResource>
      <ms:entityType>LanguageResource</ms:entityType>
      <ms:resourceName xml:lang="en">PANACEA Environment Corpus n-
↪grams EL (Greek)</ms:resourceName>
      <ms:description xml:lang="en">PANACEA Environment Corpus n-
↪grams EL (Greek) 1.0 contains Greek word n-grams and Greek word/tag/lemma n-grams.
↪in the "Environment" (ENV) domain. N-grams are accompanied by their observed
↪frequency counts. The length of the n-grams ranges from unigrams (single words) to
↪five-grams. The data were collected in the context of PANACEA (http://www.panacea-
↪lr.eu), an EU-FP7 Funded Project under Grant Agreement 248064.
The n-gram counts were generated from crawled Web pages that were automatically
↪detected to be in the Greek language and were automatically classified as relevant
↪to the ENV domain. The collection consisted of approximately 31.71 million tokens.
↪Data collection took place in the summer of 2011.</ms:description>
      <ms:version>v1.0</ms:version>
      <ms:additionalInfo>
        <ms:landingPage>http://nlp.ilsp.gr/panacea/D4.3/data/
↪201209/gms/env_el/README.txt</ms:landingPage>
      </ms:additionalInfo>
      <ms:additionalInfo>
        <ms:email>contact@someDomain.com</ms:email>
      </ms:additionalInfo>
      <ms:contact>
        <ms:Person>
          <ms:actorType>Person</ms:actorType>
          <ms:surname xml:lang="en">Prokopidis</
↪ms:surname>
          <ms:givenName xml:lang="en">Prokopis</
↪ms:givenName>
          <ms:email>contact@someDomain.com</ms:email>
        </ms:Person>
      </ms:contact>
      <ms:contact>
        <ms:Person>
          <ms:actorType>Person</ms:actorType>
          <ms:surname xml:lang="en">Papavassiliou</
↪ms:surname>
          <ms:givenName xml:lang="en">Vassilis</
↪ms:givenName>
          <ms:email>contact@someDomain.com</ms:email>
        </ms:Person>
      </ms:contact>
      <ms:keyword xml:lang="en">corpus</ms:keyword>

```

(continues on next page)

(continued from previous page)

```

<ms:domain>
  <ms:categoryLabel xml:lang="en">environment</
↪ms:categoryLabel>
</ms:domain>
<ms:resourceCreator>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="en">Institute
↪for Language and Speech Processing</ms:organizationName>
    <ms:website>http://www.ilsp.gr</ms:website>
  </ms:Organization>
</ms:resourceCreator>
<ms:creationStartDate>2011-06-01</ms:creationStartDate>
<ms:creationEndDate>2011-08-31</ms:creationEndDate>
<ms:fundingProject>
  <ms:projectName xml:lang="en">Platform for Automatic,
↪Normalized Annotation and Cost-Effective Acquisition of Language Resources for
↪Human Language </ms:projectName>
  <ms:website>http://www.panacea-lr.eu</ms:website>
</ms:fundingProject>
<ms:LRS subclass>
  <ms:LanguageDescription>
    <ms:lrType>LanguageDescription</ms:lrType>
    <ms:LanguageDescriptionSubclass>
      <ms:NGramModel>
        <ms:ldSubclassType>NGramModel
↪</ms:ldSubclassType>
        <ms:baseItem>http://w3id.org/
↪meta-share/meta-share/word</ms:baseItem>
        <ms:order>5</ms:order>
      </ms:NGramModel>
    </ms:LanguageDescriptionSubclass>
    <ms:LanguageDescriptionMediaPart>
      <ms:LanguageDescriptionTextPart>
        <ms:ldMediaType>
↪LanguageDescriptionTextPart</ms:ldMediaType>
        <ms:mediaType>http://w3id.org/
↪meta-share/meta-share/text</ms:mediaType>
        <ms:lingualityType>http://
↪w3id.org/meta-share/meta-share/monolingual</ms:lingualityType>
        <ms:language>
          <ms:languageTag>el</
↪ms:languageTag>
          <ms:languageId>el</
↪ms:languageId>
        </ms:language>
        <ms:metalanguage>
          <ms:languageTag>und</
↪ms:languageTag>
          <ms:languageId>und</
↪ms:languageId>
        </ms:metalanguage>
        <ms:creationDetails xml:lang=
↪"en">automatic web crawling, automatic language detection, data preprocessing
↪(boilerpipe filtering, lemmatization & tagging)</ms:creationDetails>
      </ms:LanguageDescriptionTextPart>
    </ms:LanguageDescriptionMediaPart>

```

(continues on next page)

(continued from previous page)

```

        <ms:DatasetDistribution>
          <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
          <ms:accessLocation>http://metashare.
↪ilsp.gr:8080/repository/download/
↪490952dc1cec11e2b545842b2b6a04d78dc202de28d5421f91752610a781175e</ms:accessLocation>
          <ms:distributionTextFeature>
            <ms:size>
              <ms:amount>435189</
↪ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/unigram</ms:sizeUnit>
            </ms:size>
            <ms:size>
              <ms:amount>3.860716E6
↪</ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/bigram</ms:sizeUnit>
            </ms:size>
            <ms:size>
              <ms:amount>9.767383E6
↪</ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/trigram</ms:sizeUnit>
            </ms:size>
            <ms:size>
              <ms:amount>1.368394E7
↪</ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/four-gram</ms:sizeUnit>
            </ms:size>
            <ms:size>
              <ms:amount>1.495402E7
↪</ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/five-gram</ms:sizeUnit>
            </ms:size>
            <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
          </ms:distributionTextFeature>
          <ms:licenceTerms>
            <ms:licenceTermsName xml:lang=
↪"en">CC-BY-SA-4.0</ms:licenceTermsName>
            <ms:licenceTermsURL>https://
↪spdx.org/licenses/CC-BY-SA-4.0.html</ms:licenceTermsURL>
          </ms:licenceTerms>
          <ms:attributionText xml:lang="en">
↪This LR has been created by Athena R.C./ILSP (www.ilsp.gr) and is licensed under a
↪CC-BY-SA licence</ms:attributionText>
        </ms:DatasetDistribution>
        <ms:personalDataIncluded>>false</
↪ms:personalDataIncluded>
        <ms:sensitiveDataIncluded>>false</
↪ms:sensitiveDataIncluded>
      </ms:LanguageDescription>
    </ms:LRSubclass>
  </ms:LanguageResource>

```

(continues on next page)

(continued from previous page)

```
</ms:DescribedEntity>
</ms:MetadataRecord>
```

9.2 Minimal version metadata for language descriptions

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources** including data language resources are presented in section *Minimal version - List of elements common to all LRTs*. **In addition**, the metadata elements that are required or recommended for language descriptions are described below.

For a quick guide to the ELG template, see *Template - Explanations*.

9.2.1 LanguageDescription

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LanguageDescription

Data type component

Optionality Mandatory

Explanation & Instructions

Wraps together elements for language descriptions

Example

```
<ms:LRSubclass>
  <ms:LanguageDescription>
    <ms:lrType>LanguageDescription</ms:lrType>
    ...
  </ms:LanguageDescription>
</ms:LRSubclass>
```

9.2.2 LanguageDescriptionSubclass

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LanguageDescription.LanguageDescriptionSubclass

Data type component

Optionality Mandatory

Explanation & Instructions

The type of the language description (used for documentation purposes)

It wraps the set of elements that must be used for the Language Description subclasses:

- Machine Learning Model: See *MLModel*
- N-gram model: See *NGramModel*
- Computational grammar: See *Grammar*

Example

```
<ms:LanguageDescriptionSubclass>
  ...
</ms:LanguageDescriptionSubclass>
```

9.2.3 MLModel

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.LanguageDescription.LanguageDescriptionSubclass.MLModel`

Data type Component

Optionality Mandatory if applicable

Explanation & Instructions

Mandatory for Machine Learning (ML) models; a ML model, for our purposes, is defined as “The model artifact that is created through a training process involving an ML algorithm (that is, the learning algorithm) and the training data to learn from”

The following set of elements are mandatory or recommended for ML models:

- `ldSubclassType` (Mandatory): Used to mark the subclass of a language description. For ML models, the value is fixed to ‘MLModel’.
- `modelVariant` (Recommended): Introduces a label that can be used to identify the variant of a ML model.
- `typesystem` (Recommended): Specifies the typesystem (preferably through an identifier or URL) that has been used for the annotation of a resource or that is required for the input resource of a tool/service or that should be used (dependency) for the annotation or used in the training of a ML model.
- `method` (Recommended): Specifies the method used for the development of a tool/service or the ML model. You must use one of the values from the CV.
- `mlFramework` (Recommended): Specifies the framework that has been used for developing a model (e.g. keras, tensorflow, etc.).
- `trainingCorpusDetails` (Recommended): Provides a detailed description of the training corpus (e.g., size, number of features , etc.).

Example

```
<ms:MLModel>
  <ms:ldSubclassType>MlModel</ms:ldSubclassType>
  <ms:modelVariant>factored</ms:modelVariant>
  <ms:typesystem>
    <ms:resourceName xml:lang="en">Universal dependencies</
↪ms:resourceName>
    <ms:version>undefined</ms:version>
  </ms:typesystem>
  <ms:method>http://w3id.org/meta-share/omtd-share/DeepLearning</ms:method>
  <ms:mlFramework>tensorflow</ms:mlFramework>
  <ms:trainingCorpusDetails xml:lang="en">Trained on a corpus of tweets</
↪ms:trainingCorpusDetails>
</ms:MLModel>
```

9.2.4 NGramModel

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LanguageDescription.LanguageDescriptionSubclass.NGramModel

Data type Component

Optionality Mandatory if applicable

Explanation & Instructions

Mandatory for n-gram models; n-gram model for our purposes is defined as “A language model consisting of n-grams, i.e. specific sequences of a number of words”

The following set of elements are mandatory or recommended for Machine Learning models:

- `ldSubclassType` (Mandatory): Used to mark the subclass of a language description. For ML models, the value is fixed to ‘NGramModel’.
- `baseItem` (Mandatory): Type of item that is represented in the n-gram resource.
- `order` (Mandatory): Specifies the maximum number of items in the sequence.
- `perplexity` (Recommended): Provides information on the perplexity derived from running on test set taken from the same corpus.

Example

```
<ms:NGramModel>
  <ms:ldSubclassType>NGramModel</ms:ldSubclassType>
  <ms:baseItem>http://w3id.org/meta-share/meta-share/word</ms:baseItem>
  <ms:order>5</ms:order>
</ms:NGramModel>
```

9.2.5 Grammar

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LanguageDescription.LanguageDescriptionSubclass.Grammar

Data type Component

Optionality Mandatory if applicable

Explanation & Instructions

Mandatory for grammars; grammar for our purposes is defined as “A set of rules governing what strings are valid or allowable in a language or text” [<https://en.oxforddictionaries.com/definition/grammar>]

The following set of elements are mandatory or recommended for computational grammars:

- `ldSubclassType` (Mandatory): Used to mark the subclass of a language description. For grammars, the value is fixed to ‘Grammar.’
- `encodingLevel` (Mandatory): Classifies the contents of a lexical/conceptual resource or language description as regards the linguistic level of analysis it caters for.
- `compliesWith` (Recommended): Specifies the vocabulary/standard/best practice to which a resource is compliant with.
- `formalism` (Recommended): Specifies the formalism (bibliographic reference, URL, name) used for the creation/enrichment of the resource (grammar or tool/service).

- `ldTask` (Recommended): Specifies the task performed by the language description.

Example

```
<ms:Grammar>
  <ms:ldSubclassType>Grammar</ms:ldSubclassType>
  <ms:encodingLevel>http://w3id.org/meta-share/meta-share/morphology</
↪ms:encodingLevel>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/GrAF</ms:compliesWith>
</ms:Grammar>
```

9.2.6 LanguageDescriptionTextPart

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.LanguageDescriptionTextPart`

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

The textual part (or whole set) of a language description

You can repeat the group of elements for multiple textual parts.

The mandatory or recommended elements for the text part of lexical/conceptual resources are:

- `mediaType` (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value ‘text’.
- `lingualityType` (Mandatory): Indicates whether the resource includes one, two or more languages.
- `multilingualityType` (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If `lingualityType` = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and multilingualSingleText (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language).
- `language` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `languageVariety` (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.
- `metalanguage` (Recommended): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).

Example

```
<ms:LanguageDescriptionMediaPart>
  <ms:LanguageDescriptionTextPart>
    <ms:ldMediaType>LanguageDescriptionTextPart</ms:ldcrMediaType>
    <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
    <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</
↪ms:lingualityType>
    <ms:language>
      <ms:languageTag>es</ms:languageTag>
```

(continues on next page)

(continued from previous page)

```

        <ms:languageId>es</ms:languageId>
      </ms:language>
      <ms:metalanguage>
        <ms:languageTag>en</ms:languageTag>
        <ms:languageId>en</ms:languageId>
      </metalanguage>
    </ms:language>
  </ms:LanguageDescriptionTextPart>
</ms:LanguageDescriptionMediaPart>

```

9.2.7 DatasetDistribution

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.DatasetDistribution

Data type component

Optionality Mandatory

Explanation & Instructions

Any form with which a dataset is distributed, such as a downloadable form in a specific format (e.g., spreadsheet, plain text, etc.) or an API with which it can be accessed

You can repeat the element for multiple distributions.

The list of mandatory and recommended elements are:

- **DatasetDistributionForm** (Mandatory): The form (medium/channel) used for distributing a language resource consisting of data (e.g., a corpus, a lexicon, etc.). The typical values are ‘downloadable’, ‘accessibleThroughInterface’, ‘accessibleThroughQuery’ (see more at [DatasetDistributionForm](#)).
- **downloadLocation** (Mandatory if applicable): A URL where the language resource (mainly data but also downloadable software programmes or forms) can be downloaded from. Use this element if the value of `datasetDistributionForm` is ‘downloadable’ and only for direct download links (i.e., from which the dataset is downloaded without the need of further actions such as clicks on a page).
- **accessLocation** (Mandatory if applicable): A URL where the resource can be accessed from; it can be used for landing pages or for cases where the resource is accessible via an interface, i.e. cases where the resource itself is not provided with a direct link for downloading. Use if the value of `datasetDistributionForm` is ‘accessibleThroughInterface’ or ‘accessibleThroughQuery’ but also for links used for downloading corpora which are mentioned on a landing page or require some kind of action on the part of the user.
- **licenceTerms** (Mandatory): See [licenceTerms](#)
- **cost** (Mandatory if applicable): Introduces the cost for accessing a resource, formally described as a set of amount and currency unit. Please use only for resources available at a cost and not for free resources.

Depending on the parts of the corpus, you must also use one or more of the following:

- **distributionTextFeature**: See [distributionTextFeature](#).

Example

```

<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/downloadable
  ↪ </ms:DatasetDistributionForm>

```

(continues on next page)

(continued from previous page)

```

    <ms:accessLocation>https://www.someAccessURL</ms:accessLocation>
    <ms:distributionTextFeature>
      <ms:size>
        <ms:amount>17601</ms:amount>
        <ms:sizeUnit>http://w3id.org/meta-share/meta-share/unit</
↪ms:sizeUnit>
      </ms:size>
      <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
↪ms:dataFormat>
      <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
    </ms:distributionTextFeature>
    <ms:licenceTerms>
      <ms:licenceTermsName xml:lang="en">openUnder-PSI</ms:licenceTermsName>
      <ms:licenceTermsURL>https://elrc-share.eu/terms/openUnderPSI.html</
↪ms:licenceTermsURL>
    </ms:licenceTerms>
  </ms:DatasetDistribution>

<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪accessibleThroughInterface</ms:DatasetDistributionForm>
  <ms:accessLocation>https://www.someAccessURL</ms:accessLocation>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>100</ms:amount>
      <ms:sizeUnit>http://w3id.org/meta-share/meta-share/text1</
↪ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf</
↪ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
    </ms:distributionTextFeature>
    <ms:licenceTerms>
      <ms:licenceTermsName xml:lang="en">some commercial licence</
↪ms:licenceTermsName>
      <ms:licenceTermsURL>https://elrc-share.eu/terms/someCommercialLicence.
↪html</ms:licenceTermsURL>
    </ms:licenceTerms>
    <ms:cost>
      <ms:amount>10000</ms:amount>
      <ms:currency>http://w3id.org/meta-share/meta-share/euro</ms:currency>
    </ms:cost>
  </ms:DatasetDistribution>

```

9.2.8 personalDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.
personalDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains personal data (mainly in the sense falling under the GDPR)

If the resource contains personal data, you can use the (optional) `personalDataDetails` to provide more information.

Example

```
<ms:personalDataIncluded>true</ms:personalDataIncluded>
<ms:personalDataDetails>The corpus contains data on the place of living and place of
↳ birth of participants</ms:personalDataDetails>
```

9.2.9 sensitiveDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.sensitiveDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains sensitive data (e.g., medical/health-related, etc.) and thus requires special handling

If the resource contains sensitive data, you can use the (optional) `sensitiveDataDetails` to provide more information.

Example

```
<ms:sensitiveDataIncluded>true</ms:sensitiveDataIncluded>
<ms:sensitiveDataDetails>The corpus contains medical data for persons with
↳ disabilities</ms:sensitiveDataDetails>
```

9.2.10 anonymized

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.anonymized

Data type boolean

Optionality Mandatory if applicable

Explanation & Instructions

Indicates whether the language resource has been anonymized

The element is mandatory if either `personalDataIncluded` or `sensitiveDataIncluded` have 'true' as value; `anonymizationDetails` must also be filled in with information on the anonymization method, etc.

Example

```
<ms:anonymized>true</ms:anonymized>
<ms:anonymizationDetails>pseudonymization performed manually</ms:anonymizationDetails>
```


Describe a lexical/conceptual resource (lexicon, glossary, ontology, etc.)

In this section you will find information on how to describe a language description (model, grammar) with the minimal metadata in order to register it in the ELG platform. If you want to find more on the ELG resource types, see [Overview](#). You will also find instructions for all data resources (technical requirements, registration instructions to the platform) in [Provide a Language Resource](#).

Examples of **lexical/conceptual resources** include

- computational lexica, that are used for computational processing, and include morphological, syntactic and semantic information;
- dictionaries in digital format,
- ontologies and controlled vocabularies,
- monolingual and multilingual terminological glossaries,
- word lists, gazetteers of place names, proper names, etc.

They typically consist of a text part, but they may also comprise audio and video files, as in the case of:

- multimedia lexica with sound recordings (e.g., pronunciation of a word) and images (e.g. pictures denoting the sense of a word),
- sign language lexica with videos.

10.1 Examples of metadata records for lexical/conceptual resources

Terminological lexicon: INTERA Corpus - the Bulgarian-English terms from the BG-EN pair

Published at: <https://live.european-language-grid.eu/catalogue/#/resource/service/lcr/694>

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns="http://w3id.org/meta-share/meta-share/" xmlns:datacite=
↪ "http://purl.org/spar/datacite/" xmlns:dc="http://www.w3.org/ns/dcat#" xmlns:ms=
↪ "http://w3id.org/meta-share/meta-share/" xmlns:omtd="http://w3id.org/meta-share/
↪ omtd-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
↪ xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../Schema/ELG-SHARE.
↪ xsd">
```

(continues on next page)

(continued from previous page)

```

<ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↪org/meta-share/meta-share/elg">value automatically assigned - leave as is</
↪ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-02-02</ms:metadataCreationDate>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>curator@somedomain.com</ms:email>
  </ms:metadataCurator>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>curator@somedomain.com</ms:email>
  </ms:metadataCreator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
  <ms:DescribedEntity>
    <ms:LanguageResource>
      <ms:entityType>LanguageResource</ms:entityType>
      <ms:resourceName xml:lang="en">INTERA Corpus - the Bulgarian-
↪English terms from the BG-EN pair</ms:resourceName>
      <ms:description xml:lang="en">The Bulgarian-English terms
↪from the BG-EN pair of the INTERA corpus; written language, domain specific (law,
↪education).</ms:description>
      <ms:version>v1.0.0 (automatically assigned)</ms:version>
      <ms:additionalInfo>
        <ms:email>contact@somedomain.com</ms:email>
      </ms:additionalInfo>
      <ms:contact>
        <ms:Person>
          <ms:actorType>Person</ms:actorType>
          <ms:surname xml:lang="en">Gavriliadou</
↪ms:surname>
          <ms:givenName xml:lang="en">Maria</
↪ms:givenName>
          <ms:email>contact@somedomain.com</ms:email>
        </ms:Person>
      </ms:contact>
      <ms:keyword xml:lang="en">lexicalconceptualresource</
↪ms:keyword>
      <ms:domain>
        <ms:categoryLabel xml:lang="en">education</
↪ms:categoryLabel>
      </ms:domain>
      <ms:domain>
        <ms:categoryLabel xml:lang="en">law</ms:categoryLabel>
      </ms:domain>
      <ms:creationStartDate>2003-01-01</ms:creationStartDate>
      <ms:creationEndDate>2004-12-31</ms:creationEndDate>
      <ms:fundingProject>
        <ms:projectName xml:lang="en">Integrated European
↪language data Repository Area</ms:projectName>
        <ms:website>http://www.elda.org/intera</ms:website>
      </ms:fundingProject>
      <ms:intendedApplication>

```

(continues on next page)

(continued from previous page)

```

        <ms:LTCClassOther>machineTranslation</ms:LTCClassOther>
      </ms:intendedApplication>
      <ms:actualUse>
        <ms:usedInApplication>
          <ms:LTCClassOther>terminologyExtraction</
↪ms:LTCClassOther>
          </ms:usedInApplication>
          <ms:actualUseDetails xml:lang="en">nlpApplications</
↪ms:actualUseDetails>
        </ms:actualUse>
        <ms:isDocumentedBy>
          <ms:title xml:lang="en">Building Multilingual
↪Terminological Resources</ms:title>
        </ms:isDocumentedBy>
        <ms:isDocumentedBy>
          <ms:title xml:lang="en">Building parallel corpora for
↪eContent professionals</ms:title>
        </ms:isDocumentedBy>
        <ms:isDocumentedBy>
          <ms:title xml:lang="en">Language resources production
↪models: the case of INTERA multilingual corpus and terminology</ms:title>
        </ms:isDocumentedBy>
        <ms:isDocumentedBy>
          <ms:title xml:lang="en">D5.2 - Report on the
↪multilingual resources production</ms:title>
          <ms:DocumentIdentifier ms:DocumentIdentifierScheme=
↪"http://purl.org/spar/datacite/url">http://www.elda.org/article176.html</
↪ms:DocumentIdentifier>
          </ms:isDocumentedBy>
          <ms:relation>
            <ms:relationType xml:lang="en">isExtractedfrom</
↪ms:relationType>
            <ms:relatedLR>
              <ms:resourceName xml:lang="en">INTERA corpus</
↪ms:resourceName>
            </ms:relatedLR>
          </ms:relation>
          <ms:LRS subclass>
            <ms:LexicalConceptualResource>
              <ms:lrType>LexicalConceptualResource</
↪ms:lrType>
              <ms:lcrSubclass>http://w3id.org/meta-share/
↪meta-share/wordlist</ms:lcrSubclass>
              <ms:encodingLevel>http://w3id.org/meta-share/
↪meta-share/morphology</ms:encodingLevel>
              <ms:LexicalConceptualResourceMediaPart>
                <ms:LexicalConceptualResourceTextPart>
                  <ms:lcrMediaType>
↪LexicalConceptualResourceTextPart</ms:lcrMediaType>
                  <ms:mediaType>http://w3id.org/
↪meta-share/meta-share/text</ms:mediaType>
                  <ms:lingualityType>http://
↪w3id.org/meta-share/meta-share/bilingual</ms:lingualityType>
                  <ms:multilingualityType>http://
↪w3id.org/meta-share/meta-share/parallel</ms:multilingualityType>
                  <ms:language>
                    <ms:languageTag>bg</
↪ms:languageTag>

```

(continues on next page)

(continued from previous page)

```

↪ms:languageId>
<ms:languageId>bg</
</ms:language>
<ms:language>
  <ms:languageTag>en</
↪ms:languageTag>
  <ms:languageId>en</
↪ms:languageId>
</ms:language>
<ms:metalinguage>
  <ms:languageTag>und</
↪ms:languageTag>
  <ms:languageId>und</
↪ms:languageId>
</ms:metalinguage>
<ms:modalityType>http://w3id.
↪org/meta-share/meta-share/writtenLanguage</ms:modalityType>
</
↪ms:LexicalConceptualResourceTextPart>
  </ms:LexicalConceptualResourceMediaPart>
  <ms:DatasetDistribution>
    <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
    <ms:accessLocation>http://metashare.
↪ilsp.gr:8080/repository/download/
↪cdba66329e8111e581e1842b2b6a04d770c91fa84de04f259240aa450aaa9081</ms:accessLocation>
    <ms:distributionTextFeature>
      <ms:size>
        <ms:amount>7581</
↪ms:amount>
        <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/word3</ms:sizeUnit>
      </ms:size>
      <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
    </ms:distributionTextFeature>
    <ms:licenceTerms>
      <ms:licenceTermsName xml:lang=
↪"en">CC-BY-4.0</ms:licenceTermsName>
      <ms:licenceTermsURL>https://
↪spdx.org/licenses/CC-BY-4.0.html</ms:licenceTermsURL>
      <ms:LicenceIdentifier
↪ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG-ENT-LIC-
↪270220-00000072</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="en">The
↪INTERA Corpus - the Bulgarian-English terms from the BG-EN pair of the ILSP/RC
↪Athena licensed under CC-BY as accessed via META-SHARE</ms:attributionText>
  </ms:DatasetDistribution>
  <ms:personalDataIncluded>>false</
↪ms:personalDataIncluded>
  <ms:sensitiveDataIncluded>>false</
↪ms:sensitiveDataIncluded>
</ms:LexicalConceptualResource>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>

```

(continues on next page)

(continued from previous page)

</ms:MetadataRecord>

Computational lexicon: MCL - Multifunctional Computational Lexicon of Contemporary PortuguesePublished at: <https://live.european-language-grid.eu/catalogue/#/resource/service/lcr/918>

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi="http://
↳ www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://w3id.org/meta-share/
↳ meta-share/ ../../Schema/ELG-SHARE.xsd">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↳ org/meta-share/meta-share/elg">value automatically assigned - leave as is</
↳ ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2005-05-12</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-02-24</ms:metadataLastDateUpdated>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>curator@somedomain.com</ms:email>
  </ms:metadataCurator>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>curator@somedomain.com</ms:email>
  </ms:metadataCreator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↳ ms:compliesWith>
  <ms:DescribedEntity>
    <ms:LanguageResource>
      <ms:entityType>LanguageResource</ms:entityType>
      <ms:resourceName xml:lang="en">MCL - Multifunctional
↳ Computational Lexicon of Contemporary Portuguese</ms:resourceName>
      <ms:description xml:lang="en">MCL is a 26,443 lemma Frequency
↳ Lexicon with 140,315 tokens, with the minimum lemma frequency of 6, extracted from
↳ CORLEX, a contemporary Portuguese corpus (16,210,438 words). CORLEX is a subcorpus
↳ of the Reference Corpus of Contemporary Portuguese and contains written and spoken
↳ texts of several types, being genre diversity a characteristic of this corpus.
↳ CORLEX contains mainly journalistic texts (56% of the written corpus and 53% of the
↳ whole corpus). In order to extract the lexicon, all the different lexical forms
↳ occurring in the corpus were indexed and subsequently tagged morphosyntactically
↳ and lemmatised by PALAVROSO. Each lemma in MCL is followed by morphosyntactic and
↳ quantitative information. The same information is given regarding each lemma token
↳ (inflected forms and some compounds). The lexicon indexations are listed in
↳ alphabetical order or decreasing frequency order.</ms:description>
      <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-
↳ share/meta-share/islrn">489-956-642-755-8</ms:LRIdentifier>
      <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-
↳ share/meta-share/other">ELRA-L0096</ms:LRIdentifier>
      <ms:version>1.0</ms:version>
      <ms:versionDate>2016-01-20</ms:versionDate>
      <ms:additionalInfo>
        <ms:landingPage>http://catalog.elra.info/product_info.
↳ php?products_id=1254</ms:landingPage>
      </ms:additionalInfo>
    </ms:LanguageResource>
  </ms:DescribedEntity>
</ms:MetadataRecord>

```

(continues on next page)

(continued from previous page)

```

<ms:additionalInfo>
  <ms:email>contact@somedomain.com</ms:email>
</ms:additionalInfo>
<ms:keyword xml:lang="en">lexicalconceptualresource</
↪ms:keyword>
  <ms:LRSubclass>
    <ms:LexicalConceptualResource>
      <ms:lrType>LexicalConceptualResource</
↪ms:lrType>
      <ms:lcrSubclass>http://w3id.org/meta-share/
↪meta-share/lexicon</ms:lcrSubclass>
      <ms:encodingLevel>http://w3id.org/meta-share/
↪meta-share/morphology</ms:encodingLevel>
      <ms:encodingLevel>http://w3id.org/meta-share/
↪meta-share/syntax</ms:encodingLevel>
      <ms:LexicalConceptualResourceMediaPart>
        <ms:LexicalConceptualResourceTextPart>
          <ms:lcrMediaType>
↪LexicalConceptualResourceTextPart</ms:lcrMediaType>
          <ms:mediaType>http://w3id.org/
↪meta-share/meta-share/text</ms:mediaType>
          <ms:lingualityType>http://
↪w3id.org/meta-share/meta-share/monolingual</ms:lingualityType>
          <ms:language>
            <ms:languageTag>pt</
↪ms:languageTag>
            <ms:languageId>pt</
↪ms:languageId>
          </ms:language>
          <ms:metalanguage>
            <ms:languageTag>und</
↪ms:languageTag>
            <ms:languageId>und</
↪ms:languageId>
          </ms:metalanguage>
        </ms:LexicalConceptualResourceMediaPart>
        <ms:DatasetDistribution>
          <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
          <ms:distributionTextFeature>
            <ms:size>
              <ms:amount>26443</
↪ms:amount>
              <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
            </ms:size>
            <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
            <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
          </ms:distributionTextFeature>
          <ms:licenceTerms>
            <ms:licenceTermsName xml:lang=
↪"en">ELRA-VAR-ACADEMIC-MEMBER-COMMERCIALUSE-1.0</ms:licenceTermsName>
            <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-VAR-ACADEMIC-MEMBER-COMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>

```

(continues on next page)

(continued from previous page)

```

</ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>

<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang=
↪"en">ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0</ms:licenceTermsName>
    <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
  </ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>

<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>

```

(continues on next page)

(continued from previous page)

```

</ms:size>
<ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
<ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
</ms:distributionTextFeature>
<ms:licenceTerms>
  <ms:licenceTermsName xml:lang=
↪"en">ELRA-VAR-COMMERCIAL-MEMBER-COMMERCIALUSE-1.0</ms:licenceTermsName>
  <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-VAR-COMMERCIAL-MEMBER-COMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
</ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>
<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang=
↪"en">ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0</ms:licenceTermsName>
    <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
  </ms:licenceTerms>
  <ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>
  <ms:distributionRightsHolder>
    <ms:Organization>
      <ms:actorType>
↪Organization</ms:actorType>
      <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
      <ms:website>http://
↪www.elra.info/en/</ms:website>

```

(continues on next page)

(continued from previous page)

```

</ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang=
↪"en">ELRA-VAR-ACADEMIC-NOMEMBER-COMMERCIALUSE-1.0</ms:licenceTermsName>
    <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-VAR-ACADEMIC-NOMEMBER-COMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
  </ms:licenceTerms>
  <ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>
  <ms:distributionRightsHolder>
    <ms:Organization>
      <ms:actorType>
↪Organization</ms:actorType>
      <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
      <ms:website>http://
↪www.elra.info/en/</ms:website>
    </ms:Organization>
  </ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang=
↪"en">ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0</ms:licenceTermsName>
    <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>

```

(continues on next page)

(continued from previous page)

```

</ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>

<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
    <ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang=
↪"en">ELRA-VAR-COMMERCIAL-NOMEMBER-COMMERCIALUSE-1.0</ms:licenceTermsName>
    <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-VAR-COMMERCIAL-NOMEMBER-COMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
  </ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>

<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://
↪w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>26443</
↪ms:amount>
      <ms:sizeUnit>http://
↪w3id.org/meta-share/meta-share/entry</ms:sizeUnit>

```

(continues on next page)

(continued from previous page)

```

</ms:size>
<ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Pdf</ms:dataFormat>
<ms:dataFormat>http://w3id.
↪org/meta-share/omtd-share/Text</ms:dataFormat>
</ms:distributionTextFeature>
<ms:licenceTerms>
  <ms:licenceTermsName xml:lang=
↪"en">ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0</ms:licenceTermsName>
  <ms:licenceTermsURL>http://
↪www.elra.info/licenses/ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0.html</
↪ms:licenceTermsURL>
</ms:licenceTerms>
<ms:availabilityStartDate>2016-01-20</
↪ms:availabilityStartDate>
<ms:distributionRightsHolder>
  <ms:Organization>
    <ms:actorType>
↪Organization</ms:actorType>
    <ms:organizationName
↪xml:lang="en">ELRA</ms:organizationName>
    <ms:website>http://
↪www.elra.info/en/</ms:website>
  </ms:Organization>
</ms:distributionRightsHolder>
</ms:DatasetDistribution>
<ms:personalDataIncluded>>false</
↪ms:personalDataIncluded>
<ms:sensitiveDataIncluded>>false</
↪ms:sensitiveDataIncluded>
</ms:LexicalConceptualResource>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

10.2 Minimal version metadata for lexical/conceptual resources

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources** including data language resources are presented in section *Minimal version - List of elements common to all LRTs*. **In addition**, the metadata elements that are required or recommended for lexical/conceptual resources are described below.

For a quick guide to the ELG template, see *Template - Explanations*.

10.2.1 LexicalConceptualResource

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LexicalConceptualResource

Data type component

Optionality Mandatory

Explanation & Instructions

Wraps together elements for lexical/conceptual resources

Example

```
<ms:LRSubclass>
  <ms:LexicalConceptualResource>
    <ms:lrType>LexicalConceptualResource</ms:lrType>
    ...
  </ms:LexicalConceptualResource>
</ms:LRSubclass>
```

10.2.2 lcrSubclass

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LexicalConceptualResource.lcrSubclass

Data type CV (lcrSubclass)

Optionality Recommended

Explanation & Instructions

Introduces a classification of lexical/conceptual resources into types (used for descriptive reasons)

Example

```
<lcrSubclass>http://w3id.org/meta-share/meta-share/computationalLexicon</lcrSubclass>
<lcrSubclass>http://w3id.org/meta-share/meta-share/ontology</lcrSubclass>
```

10.2.3 encodingLevel

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.
LexicalConceptualResource.encodingLevel

Data type CV (encodingLevel)

Optionality Mandatory

Explanation & Instructions

Classifies the contents of a lexical/conceptual resource or language description as regards the linguistic level of analysis it caters for

You can repeat the element for multiple encoding levels.

Example

```
<ms:encodingLevel>http://w3id.org/meta-share/meta-share/phonology</ms:encodingLevel>
<ms:encodingLevel>http://w3id.org/meta-share/meta-share/semantics</ms:encodingLevel>
```

10.2.4 ContentType

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.LexicalConceptualResource.ContentType

Data type CV (ContentType)

Optionality Mandatory

Explanation & Instructions

A more detailed account of the linguistic information contained in the lexical/conceptual resource

You can repeat the element for multiple encoding levels.

Example

```
<ms:ContentType>http://w3id.org/meta-share/meta-share/collocation</ms:ContentType>
<ms:ContentType>http://w3id.org/meta-share/meta-share/definition</ms:ContentType>
```

10.2.5 compliesWith

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.LexicalConceptualResource.ContentType

Data type CV (compliesWith)

Optionality Mandatory

Explanation & Instructions

Specifies the vocabulary/standard/best practice to which a resource is compliant with

You can repeat the element for multiple encoding levels.

Example

```
<ms:compliesWith>http://w3id.org/meta-share/meta-share/LMF</ms:compliesWith>
```

10.2.6 LexicalConceptualResourceTextPart

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corporus.CorporusMediaPart.LexicalConceptualResourceTextPart

Data type component

Optionality Mandatory if applicable

Explanation & Instructions

A part (or whole set) of a lexical/conceptual resource that consists of textual elements

You can repeat the group of elements for multiple textual parts.

The mandatory or recommended elements for the text part of lexical/conceptual resources are:

- `mediaType` (Mandatory): Specifies the media type of a language resource (the physical medium of the contents representation). For text parts, always use the value ‘text’.
- `lingualityType` (Mandatory): Indicates whether the resource includes one, two or more languages.
- `multilingualityType` (Mandatory if applicable): Indicates whether the resource (part) is parallel, comparable or mixed. If `lingualityType` = bilingual or multilingual, it is required; select one of the values for parallel (e.g., original text and its translations), comparable (e.g. corpus of the same domain in multiple languages) and `multilingualSingleText` (for corpora that consist of segments including text in two or more languages (e.g., the transcription of a European Parliament session with MPs speaking in their native language)).
- `language` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `languageVariety` (Mandatory if applicable): Relates a language resource that contains segments in a language variety (e.g., dialect, jargon) to it. Please use for dialect corpora.
- `metalanguage` (Mandatory): Specifies the language that is used in the resource part, expressed according to the BCP47 recommendation. See [language](#).
- `modalityType` (Recommended if applicable): Specifies the type of the modality represented in the resource. For instance, you can use ‘spoken language’ to describe transcribed speech corpora.

Example

```
<ms:LexicalConceptualResourceMediaPart>
  <ms:LexicalConceptualResourceTextPart>
    <ms:lcrMediaType>LexicalConceptualResourceTextPart</ms:lcrMediaType>
    <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
    <ms:lingualityType>http://w3id.org/meta-share/meta-share/bilingual</
↪ms:lingualityType>
    <ms:multilingualityType>http://w3id.org/meta-share/meta-share/parallel
↪</ms:multilingualityType>
    <ms:language>
      <ms:languageTag>en-US</ms:languageTag>
      <ms:languageId>en</ms:languageId>
      <ms:regionId>US</ms:regionId>
    </ms:language>
    <ms:language>
      <ms:languageTag>es</ms:languageTag>
      <ms:languageId>es</ms:languageId>
    </ms:language>
    <ms:metalanguage>
      <ms:languageTag>es</ms:languageTag>
      <ms:languageId>es</ms:languageId>
    </metalanguage>
    </ms:language>
  </ms:LexicalConceptualResourceTextPart>
</ms:LexicalConceptualResourceMediaPart>
```

10.2.7 DatasetDistribution

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpora.DatasetDistribution`

Data type component

*Optionality Mandatory**Explanation & Instructions*

Any form with which a dataset is distributed, such as a downloadable form in a specific format (e.g., spreadsheet, plain text , etc.) or an API with which it can be accessed

You can repeat the element for multiple distributions.

The list of mandatory and recommended elements are:

- `DatasetDistributionForm` (Mandatory): The form (medium/channel) used for distributing a language resource consisting of data (e.g., a corpus, a lexicon, etc.). The typical values are ‘downloadable’, ‘accessibleThroughInterface’, ‘accessibleThroughQuery’ (see more at [DatasetDistributionForm](#)).
- `downloadLocation` (Mandatory if applicable): A URL where the language resource (mainly data but also downloadable software programmes or forms) can be downloaded from. Use this element if the value of `datasetDistributionForm` is ‘downloadable’ and only for direct download links (i.e., from which the dataset is downloaded without the need of further actions such as clicks on a page).
- `accessLocation` (Mandatory if applicable): A URL where the resource can be accessed from; it can be used for landing pages or for cases where the resource is accessible via an interface, i.e. cases where the resource itself is not provided with a direct link for downloading. Use if the value of `datasetDistributionForm` is ‘accessibleThroughInterface’ or ‘accessibleThroughQuery’ but also for links used for downloading corpora which are mentioned on a landing page or require some kind of action on the part of the user.
- `licenceTerms` (Mandatory): See [licenceTerms](#).
- `cost` (Mandatory if applicable): Introduces the cost for accessing a resource, formally described as a set of amount and currency unit. Please use only for resources available at a cost and not for free resources.

Depending on the parts of the corpus, you must also use one or more of the following:

- `distributionTextFeature`: See [distributionTextFeature](#)

Example

```
<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/downloadable
</ms:DatasetDistributionForm>
  <ms:accessLocation>https://www.someAccessURL</ms:accessLocation>
  <ms:distributionTextFeature>
    <ms:size>
      <ms:amount>17601</ms:amount>
      <ms:sizeUnit>http://w3id.org/meta-share/meta-share/unit</
<ms:sizeUnit>
    </ms:size>
    <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
<ms:dataFormat>
    <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
<ms:characterEncoding>
  </ms:distributionTextFeature>
  <ms:licenceTerms>
    <ms:licenceTermsName xml:lang="en">openUnder-PSI</ms:licenceTermsName>
    <ms:licenceTermsURL>https://elrc-share.eu/terms/openUnderPSI.html</
<ms:licenceTermsURL>
  </ms:licenceTerms>
</ms:DatasetDistribution>

<ms:DatasetDistribution>
  <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
<accessibleThroughInterface</ms:DatasetDistributionForm>
```

(continues on next page)

(continued from previous page)

```

    <ms:accessLocation>https://www.someAccessURL</ms:accessLocation>
    <ms:distributionTextFeature>
      <ms:size>
        <ms:amount>100</ms:amount>
        <ms:sizeUnit>http://w3id.org/meta-share/meta-share/text1</
↪ms:sizeUnit>
      </ms:size>
      <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf</
↪ms:dataFormat>
      <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ms:characterEncoding>
    </ms:distributionTextFeature>
    <ms:licenceTerms>
      <ms:licenceTermsName xml:lang="en">some commercial licence</
↪ms:licenceTermsName>
      <ms:licenceTermsURL>https://elrc-share.eu/terms/someCommercialLicence.
↪html</ms:licenceTermsURL>
    </ms:licenceTerms>
    <ms:cost>
      <ms:amount>10000</ms:amount>
      <ms:currency>http://w3id.org/meta-share/meta-share/euro</ms:currency>
    </ms:cost>
  </ms:DatasetDistribution>

```

10.2.8 personalDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.
personalDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains personal data (mainly in the sense falling under the GDPR)

If the resource contains personal data, you can use the (optional) `personalDataDetails` to provide more information.

Example

```

<ms:personalDataIncluded>true</ms:personalDataIncluded>
<ms:personalDataDetails>The corpus contains data on the place of living and place of
↪birth of participants</ms:personalDataDetails>

```

10.2.9 sensitiveDataIncluded

Path MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.
sensitiveDataIncluded

Data type boolean

Optionality Mandatory

Explanation & Instructions

Specifies whether the language resource contains sensitive data (e.g., medical/health-related, etc.) and thus requires special handling

If the resource contains sensitive data, you can use the (optional) `sensitiveDataDetails` to provide more information.

Example

```
<ms:sensitiveDataIncluded>true</ms:sensitiveDataIncluded>
<ms:sensitiveDataDetails>The corpus contains medical data for persons with_
↪disabilities</ms:sensitiveDataDetails>
```

10.2.10 anonymized

Path `MetadataRecord.DescribedEntity.LanguageResource.LRSubclass.Corpus.anonymized`

Data type boolean

Optionality Mandatory if applicable

Explanation & Instructions

Indicates whether the language resource has been anonymized

The element is mandatory if either `personalDataIncluded` or `sensitiveDataIncluded` have 'true' as value; `anonymizationDetails` must also be filled in with information on the anonymization method, etc.

Example

```
<ms:anonymized>true</ms:anonymized>
<ms:anonymizationDetails>pseudonymization performed manually</ms:anonymizationDetails>
```


Describe a project

In this section you will find information on how to describe a project with the minimal metadata in order to register it in the ELG platform. If you want to find more on the ELG resource types, see *Overview*.

Projects listed in ELG are projects that have funded the development of LRTs or in which they have been deployed.

11.1 Examples of metadata records for projects

Example project: Bergamot – Browser-based Multilingual Translation

Published at: <https://live.european-language-grid.eu/catalogue/#/resource/projects/392>

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../
↳ Schema/ELG-SHARE.xsd" xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi=
↳ "http://www.w3.org/2001/XMLSchema-instance">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↳ org/meta-share/meta-share/elg">value automatically assigned - leave as is</
↳ ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-01-07</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-01-07</ms:metadataLastDateUpdated>
  <!-- the metadataCurator is the person responsible for editing/updating the
↳ metadata record in the ELG system and maybe different from metadata creator (for
↳ metadata records harvested from other repos, there will be no metadata creator) -->
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <!-- please add an identifier (preferably ORCID in the format below)
↳ and/or email -->
    <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://purl.org/
↳ spar/datacite/orcid">0000-0000-0000-0000</ms:PersonalIdentifier>
    <ms:email>smith@example.com</ms:email>
  </ms:metadataCurator>
```

(continues on next page)

(continued from previous page)

```

    <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
    <ms:metadataCreator>
      <ms:actorType>Person</ms:actorType>
      <ms:surname xml:lang="en">Smith</ms:surname>
      <ms:givenName xml:lang="en">John</ms:givenName>
      <!-- please add an identifier (preferably ORCID in the format below) _
↪and/or email -->
      <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://purl.org/
↪spar/datacite/orcid">0000-0000-0000-0000</ms:PersonalIdentifier>
      <ms:email>smith@example.com</ms:email>
    </ms:metadataCreator>
    <ms:DescribedEntity>
      <ms:Project>
        <ms:entityType>Project</ms:entityType>
        <ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.
↪org/meta-share/meta-share/cordis">219608</ms:ProjectIdentifier>
        <ms:projectName xml:lang="en">Browser-based Multilingual _
↪Translation</ms:projectName>
        <ms:projectShortName xml:lang="en">Bergamot</
↪ms:projectShortName>
        <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds
↪</ms:fundingType>
        <ms:funder>
          <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">European _
↪Commission</ms:organizationName>
            <ms:website>https://ec.europa.eu/info/index_en
↪</ms:website>
          </ms:Organization>
        </ms:funder>
        <ms:fundingCountry>EU</ms:fundingCountry>
        <ms:projectStartDate>2019-01-01</ms:projectStartDate>
        <ms:projectEndDate>2021-12-31</ms:projectEndDate>
        <ms:website>https://browser.mt/</ms:website>
        <ms:logo>https://ufal.mff.cuni.cz/sites/default/files/styles/
↪drupal_projects_logo_style/public/bergamot_logo.png</ms:logo>
        <ms:LTArea>
          <ms:LTCClassRecommended>http://w3id.org/meta-share/
↪omtd-share/MachineTranslation</ms:LTCClassRecommended>
          </ms:LTArea>
          <ms:LTArea>
            <ms:LTCClassOther>Browser-based Machine Translation</
↪ms:LTCClassOther>
          </ms:LTArea>
          <ms:domain>
            <ms:categoryLabel xml:lang="en">http://w3id.org/meta-
↪share/omtd-share/NewsMediaJournalismAndPublishing</ms:categoryLabel>
          </ms:domain>
          <ms:domain>
            <ms:categoryLabel xml:lang="en">General</
↪ms:categoryLabel>
          </ms:domain>
          <ms:keyword xml:lang="en">Machine translation</ms:keyword>
          <ms:keyword xml:lang="en">translation integration</ms:keyword>
          <ms:grantNumber>825303</ms:grantNumber>

```

(continues on next page)

(continued from previous page)

```

    <ms:projectSummary xml:lang="en">'The Bergamot project will
    ↳add and improve client-side machine translation in a web browser. Unlike current
    ↳cloud-based options, running directly on users' machines empowers citizens to
    ↳preserve their privacy and increases the uptake of language technologies in Europe
    ↳in various sectors that require confidentiality. Free software integrated with an
    ↳open-source web browser, such as Mozilla Firefox, will enable bottom-up adoption by
    ↳non-experts, resulting in cost savings for private and public sector users who
    ↳would otherwise procure translation or operate monolingually. To understand and
    ↳support non-expert users, our user experience work package researches their needs
    ↳and creates the user interface. Rather than simply translating text, this
    ↳interface will expose improved quality estimates, addressing the rising public
    ↳debate on algorithmic trust. Building on quality estimation research, we will
    ↳enable users to confidently generate text in a language they do not speak, enabling
    ↳cross-lingual online form filling. To improve quality overall, dynamic domain
    ↳adaptation research addresses the peculiar writing style of a website or user by
    ↳adapting translation on the fly using local information too private to upload to
    ↳the cloud. These applications require adaptation and inference to run on desktop
    ↳hardware with compact model downloads, which we address with neural network
    ↳efficiency research. Our combined research on user experience, domain adaptation,
    ↳quality estimation, outbound translation, and efficiency support a broad browser-
    ↳based innovation plan.'</ms:projectSummary>
    <ms:cost>
        <ms:amount>2999096.25</ms:amount>
        <ms:currency>http://w3id.org/meta-share/meta-share/
    ↳euro</ms:currency>
    </ms:cost>
    <ms:ecMaxContribution>
        <ms:amount>2999096.25</ms:amount>
        <ms:currency>http://w3id.org/meta-share/meta-share/
    ↳euro</ms:currency>
    </ms:ecMaxContribution>
    <ms:fundingSchemeCategory>RIA</ms:fundingSchemeCategory>
    <ms:status>SIGNED</ms:status>
    <ms:relatedCall>H2020-ICT-2018-2</ms:relatedCall>
    <ms:relatedProgramme>H2020</ms:relatedProgramme>
    <ms:relatedSubprogramme>ICT-29-2018</ms:relatedSubprogramme>
    <ms:coordinator>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">THE UNIVERSITY OF
    ↳EDINBURGH</ms:organizationName>
        <ms:website>https://www.ed.ac.uk/</ms:website>
    </ms:coordinator>
    <ms:participatingOrganization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">TARTU ULIKOOL</
    ↳ms:organizationName>
    </ms:participatingOrganization>
    <ms:participatingOrganization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">MZ DENMARK APS</
    ↳ms:organizationName>
    </ms:participatingOrganization>
    <ms:participatingOrganization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">THE UNIVERSITY OF
    ↳SHEFFIELD</ms:organizationName>
    </ms:participatingOrganization>

```

(continues on next page)

(continued from previous page)

```

        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">UNIVERZITA KARLOVA
↪</ms:organizationName>

            <ms:website>https://www.cuni.cz/</ms:website>
        </ms:participatingOrganization>
    </ms:Project>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

Example project: European Language GridPublished at: <https://live.european-language-grid.eu/catalogue/#/resource/projects/395>

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../
↪Schema/ELG-SHARE.xsd" xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi=
↪"http://www.w3.org/2001/XMLSchema-instance">
    <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
↪org/meta-share/meta-share/elg">value automatically assigned - leave as is</
↪ms:MetadataRecordIdentifier>
    <ms:metadataCreationDate>2020-01-07</ms:metadataCreationDate>
    <ms:metadataLastDateUpdated>2020-01-07</ms:metadataLastDateUpdated>
    <!-- the metadataCurator is the person responsible for editing/updating the
↪metadata record in the ELG system and maybe different from metadata creator (for
↪metadata records harvested from other repos, there will be no metadata creator) -->
    <ms:metadataCurator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Smith</ms:surname>
        <ms:givenName xml:lang="en">John</ms:givenName>
        <!-- please add an identifier (preferably ORCID in the format below)
↪and/or email -->
        <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://purl.org/
↪spar/datacite/orcid">0000-0000-0000-0000</ms:PersonalIdentifier>
        <ms:email>smith@example.com</ms:email>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
↪ms:compliesWith>
    <ms:metadataCreator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Smith</ms:surname>
        <ms:givenName xml:lang="en">John</ms:givenName>
        <!-- please add an identifier (preferably ORCID in the format below)
↪and/or email -->
        <ms:PersonalIdentifier ms:PersonalIdentifierScheme="http://purl.org/
↪spar/datacite/orcid">0000-0000-0000-0000</ms:PersonalIdentifier>
        <ms:email>smith@example.com</ms:email>
    </ms:metadataCreator>
    <ms:DescribedEntity>
        <ms:Project>
            <ms:entityType>Project</ms:entityType>
            <ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.
↪org/meta-share/meta-share/cordis">219378</ms:ProjectIdentifier>
            <ms:projectName xml:lang="en">European Language Grid</
↪ms:projectName>
            <ms:projectShortName xml:lang="en">ELG</ms:projectShortName>

```

(continues on next page)

(continued from previous page)

```

    <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds
  </ms:fundingType>
    <ms:funder>
      <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">European_
  Commission</ms:organizationName>
        <ms:website>https://ec.europa.eu/info/index_en
  </ms:website>
      </ms:Organization>
    </ms:funder>
    <ms:fundingCountry>EU</ms:fundingCountry>
    <ms:projectStartDate>2019-01-01</ms:projectStartDate>
    <ms:projectEndDate>2021-12-31</ms:projectEndDate>
    <ms:website>https://www.european-language-grid.eu/</
  ms:website>
    <ms:logo>https://www.european-language-grid.eu/wp-content/
  themes/elg_theme/fab/image/logo/rgb_elg__logo--colour.svg</ms:logo>
    <ms:LTArea>
      <ms:LTClassRecommended>http://w3id.org/meta-share/
  omt-d-share/LanguageTechnology</ms:LTClassRecommended>
    </ms:LTArea>
    <ms:keyword xml:lang="en">Language technology services</
  ms:keyword>
    <ms:keyword xml:lang="en">Multilingualism</ms:keyword>
    <ms:keyword xml:lang="en">Less-resourced languages</
  ms:keyword>
    <ms:grantNumber>825627</ms:grantNumber>
    <ms:projectSummary xml:lang="en">With 24 official EU and many_
  more additional languages, multilingualism in Europe and an inclusive Digital_
  Single Market can only be enabled through Language Technologies (LTs). European LT_
  business is dominated by thousands of SMEs and a few large players. Many are world-
  class, with technologies that outperform the global players. However, European LT_
  business is also fragmented by nation states, languages, verticals and sectors.
  Likewise, while much of European LT research is world-class, with results_
  transferred into industry and commercial products, its full impact is held back by_
  fragmentation. The key issue and challenge is the fragmentation of the European LT_
  landscape. The European Language Grid (ELG) project will address this fragmentation_
  by establishing the ELG as the primary platform for LT in Europe. The ELG will be a_
  scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds_
  of commercial and non-commercial Language Technologies for all European languages,_
  including running tools and services as well as data sets and resources. It will_
  enable the commercial and non-commercial European LT community to deposit and_
  upload their technologies and data sets into the ELG, to deploy them through the_
  grid, and to connect with other resources. The ELG will boost the Multilingual_
  Digital Single Market towards a thriving European LT community, creating new jobs_
  and opportunities. Through open calls, up to 20 pilot projects will be financially_
  supported to demonstrate the usefulness of the ELG. The proposal is rooted in the_
  experience of a consortium with partners involved in all relevant initiatives.
  Based on these, 30\ national competence centres and the European LT Board will be_
  set up for European coordination. The ELG will foster language technologies for_
  Europe built in Europe, tailored to our languages and cultures and to our_
  societal and economical demands, benefitting the European citizen, society,_
  innovation and industry.</ms:projectSummary>
    <ms:cost>
      <ms:amount>7460206.25</ms:amount>
      <ms:currency>http://w3id.org/meta-share/meta-share/
  euro</ms:currency>

```

(continues on next page)

(continued from previous page)

```

        </ms:cost>
        <ms:ecMaxContribution>
            <ms:amount>6999631.25</ms:amount>
            <ms:currency>http://w3id.org/meta-share/meta-share/
↪euro</ms:currency>
        </ms:ecMaxContribution>
        <ms:fundingSchemeCategory>IA</ms:fundingSchemeCategory>
        <ms:status>SIGNED</ms:status>
        <ms:relatedCall>H2020-ICT-2018-2</ms:relatedCall>
        <ms:relatedProgramme>H2020</ms:relatedProgramme>
        <ms:relatedSubprogramme>ICT-29-2018</ms:relatedSubprogramme>
        <ms:coordinator>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">DEUTSCHES
↪FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH</ms:organizationName>
            <ms:website>https://www.dfki.de/</ms:website>
        </ms:coordinator>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">SAIL LABS
↪TECHNOLOGY GMBH</ms:organizationName>
            <ms:website>https://www.sail-labs.com/</ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">THE UNIVERSITY OF
↪SHEFFIELD</ms:organizationName>
            <ms:website>https://www.dfki.de/</ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">ATHINA-EREVNITIKO
↪KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS
↪GNOSIS</ms:organizationName>
            <ms:website>https://www.athena-innovation.gr/</
↪ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">EVALUATIONS AND
↪LANGUAGE RESOURCES DISTRIBUTION AGENCY</ms:organizationName>
            <ms:website>http://www.elda.org/</ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">TILDE SIA</
↪ms:organizationName>
            <ms:website>https://www.tilde.eu/</ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">UNIVERZITA KARLOVA
↪</ms:organizationName>
            <ms:website>https://www.cuni.cz/</ms:website>
        </ms:participatingOrganization>
        <ms:participatingOrganization>
            <ms:actorType>Organization</ms:actorType>

```

(continues on next page)

(continued from previous page)

```

        <ms:organizationName xml:lang="en">THE UNIVERSITY OF
↪EDINBURGH</ms:organizationName>
        <ms:website>https://www.ed.ac.uk/</ms:website>
    </ms:participatingOrganization>
    <ms:participatingOrganization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">EXPERT SYSTEM
↪IBERIA SL</ms:organizationName>
        <ms:website>http://www.expertsystem.com/</ms:website>
    </ms:participatingOrganization>
    </ms:Project>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

11.2 Minimal version metadata for projects

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources** are presented in section *Minimal version - List of elements common to all LRTs*. **In addition**, the metadata elements that are required or recommended for projects are described below.

For a quick guide to the ELG template, see *Template - Explanations*.

11.2.1 Project

Path MetadataRecord.DescribedEntity.Project

Data type component

Optionality Mandatory

Explanation & Instructions

Wraps together elements for projects

Example

```

<ms:Project>
    <ms:entityType>project</ms:entityType>
    ...
</ms:Project>

```

11.2.2 ProjectIdentifier

Path MetadataRecord.DescribedEntity.Project.ProjectIdentifier

Data type string

Optionality Recommended

Explanations & Instructions

A string (e.g., PID, internal to an organization, issued by the funding authority, etc.) used to uniquely identify a project

You must also use the attribute `ProjectIdentifierScheme` to specify the name of the scheme according to which an identifier is assigned to a project by the authority that issues it. See https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Attribute_ms_ProjectIdentifierScheme.html#ProjectIdentifierScheme for details.

Example

```
<ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.org/meta-share/meta-
↪share/cordis">219608</ms:ProjectIdentifier>

<ms:ProjectIdentifier ms:ProjectIdentifierScheme="http://w3id.org/meta-share/meta-
↪share/cordis">219378</ms:ProjectIdentifier>
```

11.2.3 projectName

Path `MetadataRecord.DescribedEntity.Project.projectName`

Data type multilingual string

Optionality Mandatory

Explanations & Instructions

The full name (title) of a project

Example

```
<ms:projectName xml:lang="en">Browser-based Multilingual Translation</ms:projectName>

<ms:projectName xml:lang="en">European Language Grid</ms:projectName>
```

11.2.4 projectShortName

Path `MetadataRecord.DescribedEntity.Project.projectShortName`

Data type multilingual string

Optionality Recommended

Explanations & Instructions

Introduces a short name (e.g., acronym, abbreviated form) by which a project is known

Example

```
<ms:projectShortName xml:lang="en">Bergamot</ms:projectShortName>

<ms:projectShortName xml:lang="en">ELG</ms:projectShortName>
```

11.2.5 projectAlternativeName

Path MetadataRecord.DescribedEntity.Project.

Data type multilingual string

Optionality Recommended

Explanations & Instructions

Introduces an alternative name (other than the short name) used for a project

Example

```
<ms:projectAlternativeName xml:lang="en">The European Language Grid</ms:projectName>
```

11.2.6 fundingType

Path MetadataRecord.DescribedEntity.Project.

Data type CV (fundingType)

Optionality Recommended

Explanations & Instructions

Specifies the type of funding of a project with regard to the source of the funding

Example

```
<ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds</ms:fundingType>
```

11.2.7 funder

Path MetadataRecord.DescribedEntity.Project.funder

Data type component

Optionality Recommended

Explanations & Instructions

Identifies the person/organization/group that has financed the project

Funding information is important for acknowledgement purposes.

For organizations, you must provide the name of the organization (organizationName) and, if possible, a website (website) and/or an identifier (OrganizationIdentifier).

Example

```
<ms:funder>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="en">European Commission</
  <ms:organizationName>
    <ms:website>https://ec.europa.eu/info/index_en</ms:website>
```

(continues on next page)

(continued from previous page)

```
</ms:Organization>
</ms:funder>
```

11.2.8 fundingCountry

Path MetadataRecord.DescribedEntity.Project.fundingCountry

Data type CV (regionIdType)

Optionality Recommended

Explanations & Instructions

Specifies the name of the funding country, in case of national funding as mentioned in ISO3166

Example

```
<ms:fundingCountry>EU</ms:fundingCountry>
```

11.2.9 website

Path MetadataRecord.DescribedEntity.Project.website

Data type URL

Optionality Recommended

Explanations & Instructions

Links to a URL that acts as the primary page (like a table of contents) introducing information about an organization (e.g., products, contact information, etc.) or project

Example

```
<ms:website>https://browser.mt/</ms:website>
<ms:website>https://www.european-language-grid.eu/</ms:website>
```

11.2.10 logo

Path MetadataRecord.DescribedEntity.Project.logo

Data type URL

Optionality Recommended

Explanations & Instructions

Links to a URL with an image file containing a symbol or graphic object used to identify the entity

Example


```
<ms:logo>https://ufal.mff.cuni.cz/sites/default/files/styles/drupal_projects_logo_
↪style/public/bergamot_logo.png</ms:logo>

<ms:logo>https://www.european-language-grid.eu/wp-content/themes/elg_theme/fab/image/
↪logo/rgb_elg__logo--colour.svg</ms:logo>
```

11.2.11 LTArea

Path MetadataRecord.DescribedEntity.Project.LTArea

Data type component

Optionality Recommended

Explanations & Instructions

Introduces a Language Technology-related area that the project deals with

For details, see https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Element_ms_LTArea.html#LTArea.

Example

```
<ms:LTArea>
  <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
  ↪MachineTranslation</ms:LTCClassRecommended>
</ms:LTArea>
<ms:LTArea>
  <ms:LTCClassOther>Browser-based Machine Translation</ms:LTCClassOther>
</ms:LTArea>
```

11.2.12 domain

Path MetadataRecord.DescribedEntity.Project.domain

Data type component

Optionality Recommended

Explanations & Instructions

Identifies a domain that the project deals with

You must fill in the CategoryLabel element with a free text value. If you prefer to add a value from an established controlled vocabulary, you can also use the DomainIdentifier (with the attribute DomainClassificationScheme with the appropriate value).

Example

```
<ms:domain>
  <ms:categoryLabel xml:lang="en">http://w3id.org/meta-share/omtd-share/
  ↪NewsMediaJournalismAndPublishing</ms:categoryLabel>
</ms:domain>
<ms:domain>
  <ms:categoryLabel xml:lang="en">General</ms:categoryLabel>
</ms:domain>
```

11.2.13 keyword

Path `MetadataRecord.DescribedEntity.Project.keyword`

Data type multilingual string

Optionality Recommended

Explanations & Instructions

Introduces a word or phrase considered important for the description of the project and thus used to index or classify it

Example

```
<ms:keyword xml:lang="en">Machine translation</ms:keyword>
<ms:keyword xml:lang="en">translation integration</ms:keyword>

<ms:keyword xml:lang="en">Language technology services</ms:keyword>
<ms:keyword xml:lang="en">Multilingualism</ms:keyword>
<ms:keyword xml:lang="en">Less-resourced languages</ms:keyword>
```

11.2.14 socialMediaOccupationalAccount

Path `MetadataRecord.DescribedEntity.Project.socialMediaOccupationalAccount`

Data type multilingual string

Optionality Recommended

Explanations & Instructions

Introduces the social media or occupational account details of a person, organization or project

You must also use the attribute `socialMediaAccountType` to specify the type of social media account. See https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Attribute_ms_socialMediaOccupationalAccountType.html#socialMediaOccupationalAccountType for details.

Example

Note: TODO: add example

11.2.15 projectSummary

Path `MetadataRecord.DescribedEntity.Project.projectSummary`

Data type multilingual string

Optionality Recommended

Explanations & Instructions

Introduces a short description (in free text) of the main objectives, mission or contents of the project

Example

```

<ms:projectSummary xml:lang="en">'The Bergamot project will add and improve client-
→side machine translation in a web browser. Unlike current cloud-based options,
→running directly on users'' machines empowers citizens to preserve their privacy,
→and increases the uptake of language technologies in Europe in various sectors that
→require confidentiality. Free software integrated with an open-source web browser,
→such as Mozilla Firefox, will enable bottom-up adoption by non-experts, resulting
→in cost savings for private and public sector users who would otherwise procure
→translation or operate monolingually. To understand and support non-expert users,
→our user experience work package researches their needs and creates the user
→interface. Rather than simply translating text, this interface will expose
→improved quality estimates, addressing the rising public debate on algorithmic
→trust. Building on quality estimation research, we will enable users to
→confidently generate text in a language they do not speak, enabling cross-lingual
→online form filling. To improve quality overall, dynamic domain adaptation
→research addresses the peculiar writing style of a website or user by adapting
→translation on the fly using local information too private to upload to the cloud.
→These applications require adaptation and inference to run on desktop hardware with
→compact model downloads, which we address with neural network efficiency research.
→Our combined research on user experience, domain adaptation, quality estimation,
→outbound translation, and efficiency support a broad browser-based innovation plan.'
</ms:projectSummary>

<ms:projectSummary xml:lang="en">With 24 official EU and many more additional
→languages, multilingualism in Europe and an inclusive Digital Single Market can
→only be enabled through Language Technologies (LTs). European LT business is
→dominated by thousands of SMEs and a few large players. Many are world-class, with
→technologies that outperform the global players. However, European LT business is
→also fragmented by nation states, languages, verticals and sectors. Likewise,
→while much of European LT research is world-class, with results transferred into
→industry and commercial products, its full impact is held back by fragmentation.
→The key issue and challenge is the fragmentation of the European LT landscape. The
→European Language Grid (ELG) project will address this fragmentation by
→establishing the ELG as the primary platform for LT in Europe. The ELG will be a
→scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds
→of commercial and non-commercial Language Technologies for all European languages,
→including running tools and services as well as data sets and resources. It will
→enable the commercial and non-commercial European LT community to deposit and
→upload their technologies and data sets into the ELG, to deploy them through the
→grid, and to connect with other resources. The ELG will boost the Multilingual
→Digital Single Market towards a thriving European LT community, creating new jobs
→and opportunities. Through open calls, up to 20 pilot projects will be financially
→supported to demonstrate the usefulness of the ELG. The proposal is rooted in the
→experience of a consortium with partners involved in all relevant initiatives.
→Based on these, 30\\ national competence centres and the European LT Board will be
→set up for European coordination. The ELG will foster language technologies for
→Europe built in Europe, tailored to our languages and cultures and to our
→societal and economical demands, benefitting the European citizen, society,
→innovation and industry.</ms:projectSummary>

```


Describe an organization

In this section you will find information on how to describe an organization with the minimal metadata in order to register in the ELG platform. If you want to find more on the ELG resource types, see [Overview](#).

Organizations listed in ELG are organizations which are or have been active in Language Technology in Europe.

12.1 Examples of metadata records for organizations

University: Charles University, Prague

```
<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../../Schema/ELG-SHARE.xsd" xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">value automatically assigned - leave as is</ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-01-07</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-01-07</ms:metadataLastDateUpdated>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>smith@example.com</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>smith@example.com</ms:email>
  </ms:metadataCreator>
```

(continues on next page)

(continued from previous page)

```

<ms:DescribedEntity>
  <ms:Organization>
    <ms:entityType>Organization</ms:entityType>
    <ms:OrganizationIdentifier ms:OrganizationIdentifierScheme="http://w3id.org/
↪meta-share/meta-share/eltg">automatically assigned by ELG - please don't change</
↪ms:OrganizationIdentifier>
      <ms:organizationName xml:lang="en">Charles University</ms:organizationName>
      <ms:organizationShortName xml:lang="en">CUNI</ms:organizationShortName>
      <ms:organizationAlternativeName xml:lang="en">UNIVERZITA KARLOVA</
↪ms:organizationAlternativeName>
      <ms:organizationRole>http://w3id.org/meta-share/meta-share/LTSupplier</
↪ms:organizationRole>
      <ms:organizationRole>http://w3id.org/meta-share/meta-share/
↪researchOrganization</ms:organizationRole>
      <ms:organizationRole>http://w3id.org/meta-share/meta-share/
↪languageServiceProvider</ms:organizationRole>
      <ms:organizationLegalStatus>http://w3id.org/meta-share/meta-share/
↪academicInstitution</ms:organizationLegalStatus>
      <ms:countryOfRegistration>CZ</ms:countryOfRegistration>
      <ms:organizationBio xml:lang="en">Charles University was founded in 1348,
↪making it one of the oldest universities in the world. Yet it is also renowned as a
↪modern, dynamic, cosmopolitan and prestigious institution of higher education. It
↪is the largest and most renowned Czech university, and is also the best-rated Czech
↪university according to international rankings. There are currently 17 faculties at
↪the University (14 in Prague, 2 in Hradec Králové and 1 in Plzeň), plus 3
↪institutes, 6 other centres of teaching, research, development and other creative
↪activities, a centre providing information services, 5 facilities serving the whole
↪University, and the Rectorate - which is the executive management body for the
↪whole University.</ms:organizationBio>
      <ms:logo>https://cuni.cz/UKEN-1-version1-afoto.jpg</ms:logo>
      <ms:LTArea>
        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
        </ms:LTArea>
        <ms:LTArea>
          <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪MachineTranslation</ms:LTClassRecommended>
          </ms:LTArea>
          <ms:LTArea>
            <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪SpeechRecognition</ms:LTClassRecommended>
            </ms:LTArea>
            <ms:LTArea>
              <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪Annotation</ms:LTClassRecommended>
              </ms:LTArea>
              <ms:LTArea>
                <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LexiconCreation</ms:LTClassRecommended>
                </ms:LTArea>
                <ms:LTArea>
                  <ms:LTClassOther>language resources creation</ms:LTClassOther>
                  </ms:LTArea>
                  <ms:LTArea>
                    <ms:LTClassOther>Dialog systems</ms:LTClassOther>
                    </ms:LTArea>
                    <ms:keyword xml:lang="en">Computational Linguistics</ms:keyword>

```

(continues on next page)

(continued from previous page)

```

<ms:keyword xml:lang="en">Natural Language Processing</ms:keyword>
<ms:keyword xml:lang="en">Language Resources</ms:keyword>
<ms:keyword xml:lang="en">Research infrastructures</ms:keyword>
<ms:keyword xml:lang="en">Language Resources</ms:keyword>
<ms:keyword xml:lang="en">Digital Humanities</ms:keyword>
<ms:website>https://www.cuni.cz</ms:website>
<ms:addressSet>
  <ms:address xml:lang="en">OVOCNY TRH 560/5</ms:address>
  <ms:zipCode>116 36</ms:zipCode>
  <ms:city xml:lang="en">PRAHA 1</ms:city>
  <ms:country>CZ</ms:country>
</ms:addressSet>
<ms:hasDivision>
  <ms:divisionName xml:lang="en">Institute of Formal and Applied
↪Linguistics</ms:divisionName>
  <ms:divisionShortName xml:lang="en">UFAL</ms:divisionShortName>
  <ms:divisionCategory>http://w3id.org/meta-share/meta-share/institute</
↪ms:divisionCategory>
  <ms:organizationBio xml:lang="en">'Institute of Formal and Applied
↪Linguistics (UFAL) at the Computer Science School, Faculty of Mathematics and
↪Physics, Charles University, Czech Republic. The institute was established in 1990
↪as a continuation of the research and teaching activities carried out by the
↪former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of
↪Philosophy and later at the Faculty of Mathematics and Physics, Charles University.
↪The Institute is a primarily research department working on many topics in the area
↪of Computational Linguistics, and on many research projects both nationally and
↪internationally. However, the Institute of Formal and Applied Linguistics is also a
↪regular department in the sense that it carries a comprehensive teaching program
↪both for the Master's degree (Mgr., or MSc.) as well as for a doctorate (Ph.D.) in
↪Computational Linguistics. Both programs are taught in Czech and English. The
↪Institute is also a member of the double- degree \"Master's LCT programme\" of the
↪EU. Students also can take advantage of the Erasmus program for typically semester-
↪long stays at partner Universities abroad. '</ms:organizationBio>
  <ms:logo>https://ufal.mff.cuni.cz/sites/all/themes/drufal/css/logo/
↪logo_ufal_110u.png</ms:logo>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
↪MachineTranslation</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
↪SpeechRecognition</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
↪Annotation</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LexiconCreation</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Lexical Resources</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Dialog systems</ms:LTCClassOther>
  </ms:LTArea>

```

(continues on next page)

(continued from previous page)

```

<ms:LTArea>
  <ms:LTClassOther>Corpus Creation</ms:LTClassOther>
</ms:LTArea>
<ms:LTArea>
  <ms:LTClassOther>Research Infrastructure</ms:LTClassOther>
</ms:LTArea>
<ms:LTArea>
  <ms:LTClassOther>LT services</ms:LTClassOther>
</ms:LTArea>
<ms:LTArea>
  <ms:LTClassOther>NLP Support</ms:LTClassOther>
</ms:LTArea>
<ms:LTArea>
  <ms:LTClassOther>Digital Humanities</ms:LTClassOther>
</ms:LTArea>
<ms:keyword xml:lang="en">Computational Linguistics</ms:keyword>
  <ms:addressSet>
    <ms:address xml:lang="en">Malostranská n. 25</ms:address>
    <ms:zipCode>11800</ms:zipCode>
    <ms:city xml:lang="en">Praha 1</ms:city>
    <ms:country>CZ</ms:country>
  </ms:addressSet>
</ms:hasDivision>
</ms:Organization>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

SME: Evaluation and Language Resources Distribution Agency (ELDA)

```

<?xml version="1.0" encoding="UTF-8"?>
<ms:MetadataRecord xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ ../.. /
Schema/ELG-SHARE.xsd" xmlns:ms="http://w3id.org/meta-share/meta-share/" xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance">
  <ms:MetadataRecordIdentifier ms:MetadataRecordIdentifierScheme="http://w3id.
org/meta-share/meta-share/elg">value automatically assigned - leave as is</
ms:MetadataRecordIdentifier>
  <ms:metadataCreationDate>2020-01-07</ms:metadataCreationDate>
  <ms:metadataLastDateUpdated>2020-01-07</ms:metadataLastDateUpdated>
  <ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>smith@example.com</ms:email>
  </ms:metadataCurator>
  <ms:compliesWith>http://w3id.org/meta-share/meta-share/ELG-SHARE</
ms:compliesWith>
  <ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Smith</ms:surname>
    <ms:givenName xml:lang="en">John</ms:givenName>
    <ms:email>smith@example.com</ms:email>
  </ms:metadataCreator>
  <ms:DescribedEntity>
    <ms:Organization>
      <ms:entityType>Organization</ms:entityType>

```

(continues on next page)

(continued from previous page)

```

    <ms:OrganizationIdentifier ms:OrganizationIdentifierScheme="http://w3id.org/
↳meta-share/meta-share/elg">automatically assigned by ELG - please don't change</
↳ms:OrganizationIdentifier>
    <ms:organizationName xml:lang="en">Evaluation and Language Resources_
↳Distribution Agency</ms:organizationName>
    <ms:organizationShortName xml:lang="en">ELDA</ms:organizationShortName>
    <ms:organizationAlternativeName xml:lang="en">EVALUATIONS AND LANGUAGE_
↳RESOURCES DISTRIBUTION AGENCY</ms:organizationAlternativeName>
    <ms:organizationRole>http://w3id.org/meta-share/meta-share/LTSupplier</
↳ms:organizationRole>
    <ms:organizationRole>http://w3id.org/meta-share/meta-share/dataEvaluator</
↳ms:organizationRole>
    <ms:organizationLegalStatus>http://w3id.org/meta-share/meta-share/sme</
↳ms:organizationLegalStatus>
    <ms:countryOfRegistration>FR</ms:countryOfRegistration>
    <ms:organizationBio xml:lang="en">The Evaluations and Language Resources_
↳Distribution Agency (ELDA), was created in 1995 as the organizational_
↳infrastructure with the mission of providing a central clearing house for Language_
↳Resources (LR) of the European Language Resources Association (ELRA). ELDA was set_
↳up to identify, classify, collect, validate and distribute the language resources_
↳that are needed by the Human Language Technology (HLT) community. Anticipating the_
↳evolutions in the HLT field, ELDA broadened its activities to cover multimedia/_
↳multimodal resources as well as evaluation activities, distributing the language_
↳resources needed for evaluation purposes, and conducting/coordinating evaluation_
↳campaigns. ELDA has played a significant role within the major Multimedia and_
↳Multimodal production projects that resulted in one of the most impressive_
↳catalogues of available data sets, embracing all aspects of Language Technologies._
↳ELDA was also involved in evaluation initiatives, in several FPs' projects_
↳involving HLT infrastructures, as well as in national programmes. In addition to_
↳work on data production, processing and annotation, validation and quality control,_
↳several of these projects also involved work on legal framework management for the_
↳produced resources. Moreover, ELDA has contributed to the development of open_
↳platforms and has joined forces with other European key players by bringing its_
↳assets (LR catalogue, evaluation services and benchmarking) to constitute Europe's_
↳backbone for Language Resources sharing and distribution. ELDA is also the_
↳initiator of the Language Resource and the Evaluation Conference (LREC), since 1998.
↳ With over 1200 participants, LREC is the major event on Language Resources (LRs)_
↳and Evaluation for Human Language Technologies (HLT).</ms:organizationBio>
    <ms:logo>https://www.european-language-grid.eu/wp-content/uploads/2019/03/
↳logo__consortium-elda.svg</ms:logo>
    <ms:LTArea>
        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↳LanguageTechnology</ms:LTClassRecommended>
        </ms:LTArea>
        <ms:LTArea>
            <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↳Evaluation</ms:LTClassRecommended>
            </ms:LTArea>
            <ms:LTArea>
                <ms:LTClassOther>Language Resource collection, processing, production
↳</ms:LTClassOther>
                </ms:LTArea>
                <ms:LTArea>
                    <ms:LTClassOther>legal clearing</ms:LTClassOther>
                    </ms:LTArea>
                    <ms:LTArea>
                        <ms:LTClassOther>HLT evaluation and dissemination</ms:LTClassOther>

```

(continues on next page)

(continued from previous page)

```
</ms:LTArea>
<ms:keyword xml:lang="en">Language Resources and Evaluation</ms:keyword>
<ms:keyword xml:lang="en">Legal support</ms:keyword>
<ms:keyword xml:lang="en">Data management</ms:keyword>
<ms:website>http://www.elra.info/en/</ms:website>
<ms:addressSet>
  <ms:address xml:lang="en">9 RUE DES CORDELIERES</ms:address>
  <ms:zipCode>75 013</ms:zipCode>
  <ms:city xml:lang="en">Paris</ms:city>
  <ms:country>FR</ms:country>
</ms:addressSet>
</ms:Organization>
</ms:DescribedEntity>
</ms:MetadataRecord>
```

12.2 Minimal version metadata for organizations

The set of the metadata (mandatory or recommended) that **are common to all kinds of resources** are presented in section *Minimal version - List of elements common to all LRTs*. **In addition**, the metadata elements that are required or recommended for projects are described below.

For a quick guide to the ELG template, see *Template - Explanations*.

12.2.1 Organization

Path MetadataRecord.DescribedEntity.Organization

Data type component

Optionality Mandatory

Explanation & Instructions

Wraps together elements for organizations

Example

```
<ms:Organization>
  <ms:entityType>organization</ms:entityType>
  ...
</ms:Organization>
```

12.2.2 OrganizationIdentifier (M)

Path MetadataRecord.DescribedEntity.Organization.OrganizationIdentifier

Data type string

Optionality Recommended

Explanation & Instructions

A string (e.g., PID, internal to an organization, issued by the funding authority, etc.) used to uniquely identify an organization

You must also use the attribute `OrganizationIdentifierScheme` to specify the name of the scheme according to which an identifier is assigned to an organization by the authority that issues it. See https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Attribute_ms_OrganizationIdentifierScheme.html#OrganizationIdentifierScheme for details.

Example

```
<ms:OrganizationIdentifier ms:OrganizationIdentifierScheme="http://w3id.org/meta-
↪share/meta-share/elg">automatically assigned by ELG - please don't change</
↪ms:OrganizationIdentifier>
```

12.2.3 organizationName (M)

Path `MetadataRecord.DescribedEntity.Organization.organizationName`

Data type multilingual string

Optionality Mandatory

Explanation & Instructions

The full name of an organization

Example

```
<ms:organizationName xml:lang="en">Charles University</ms:organizationName>

<ms:organizationName xml:lang="en">Evaluation and Language Resources Distribution_
↪Agency</ms:organizationName>
```

12.2.4 organizationShortName (M)

Path `MetadataRecord.DescribedEntity.Organization.organizationShortName`

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces the short name (abbreviation, acronym , etc.) used for an organization

Example

```
<ms:organizationShortName xml:lang="en">CUNI</ms:organizationName>

<ms:organizationShortName xml:lang="en">ELDA</ms:organizationName>
```

12.2.5 organizationAlternativeName (M)

Path MetadataRecord.DescribedEntity.Organization.organizationAlternativeName

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces an alternative name (other than the short name) used for an organization

Example

```
<ms:organizationAlternativeName xml:lang="en">UNIVERZITA KARLOVA</ms:organizationAlternativeName>

<ms:organizationAlternativeName xml:lang="en">EVALUATIONS AND LANGUAGE RESOURCES_
DISTRIBUTION AGENCY</ms:organizationAlternativeName>
```

12.2.6 countryOfRegistration (M)

Path MetadataRecord.DescribedEntity.Organization.countryOfRegistration

Data type CV ([regionIdType](#))

Optionality Recommended

Explanation & Instructions

Introduces the country in which an organization has been first registered as a legal entity

Example

```
<ms:countryOfRegistration>CZ</ms:countryOfRegistration>

<ms:countryOfRegistration>FR</ms:countryOfRegistration>
```

12.2.7 organizationBio (M)

Path MetadataRecord.DescribedEntity.Organization.organizationBio

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces a short free-text account that provides information on an organization

Example

```
<ms:organizationBio xml:lang="en">Charles University was founded in 1348, making it
→one of the oldest universities in the world. Yet it is also renowned as a modern,
→dynamic, cosmopolitan and prestigious institution of higher education. It is the
→largest and most renowned Czech university, and is also the best-rated Czech
→university according to international rankings. There are currently 17 faculties at
→the University (14 in Prague, 2 in Hradec Králové and 1 in Plzeň), plus 3
→institutes, 6 other centres of teaching, research, development and other creative
→activities, a centre providing information services, 5 facilities serving the whole
→University, and the Rectorate – which is the executive management body for the
→whole University.</ms:organizationBio>
```

```
<ms:organizationBio xml:lang="en">The Evaluations and Language Resources Distribution
→Agency (ELDA), was created in 1995 as the organizational infrastructure with the
→mission of providing a central clearing house for Language Resources (LR) of the
→European Language Resources Association (ELRA). ELDA was set up to identify,
→classify, collect, validate and distribute the language resources that are needed
→by the Human Language Technology (HLT) community. Anticipating the evolutions in
→the HLT field, ELDA broadened its activities to cover multimedia/multimodal
→resources as well as evaluation activities, distributing the language resources
→needed for evaluation purposes, and conducting/coordinating evaluation campaigns.
→ELDA has played a significant role within the major Multimedia and Multimodal
→production projects that resulted in one of the most impressive catalogues of
→available data sets, embracing all aspects of Language Technologies. ELDA was also
→involved in evaluation initiatives, in several FPs' projects involving HLT
→infrastructures, as well as in national programmes. In addition to work on data
→production, processing and annotation, validation and quality control, several of
→these projects also involved work on legal framework management for the produced
→resources. Moreover, ELDA has contributed to the development of open platforms and
→has joined forces with other European key players by bringing its assets (LR
→catalogue, evaluation services and benchmarking) to constitute Europe's backbone
→for Language Resources sharing and distribution. ELDA is also the initiator of the
→Language Resource and the Evaluation Conference (LREC), since 1998. With over 1200
→participants, LREC is the major event on Language Resources (LRs) and Evaluation
→for Human Language Technologies (HLT).</ms:organizationBio>
```

12.2.8 logo

Path MetadataRecord.DescribedEntity.Organization.logo

Data type URL

Optionality Recommended

Explanation & Instructions

Links to a URL with an image file containing a symbol or graphic object used to identify the entity

Example

```
<ms:logo>https://cuni.cz/UKEN-1-version1-afoto.jpg</ms:logo>
```

```
<ms:logo>https://www.european-language-grid.eu/wp-content/uploads/2019/03/logo__
→consortium-elda.svg</ms:logo>
```

12.2.9 LTArea (M)

Path MetadataRecord.DescribedEntity.Organization.LTArea

Data type component

Optionality Recommended

Explanation & Instructions

Introduces a Language Technology-related area that a person or organization is involved or active in

For details, see https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Element_ms_LTArea.html#LTArea.

Example

```
<ms:LTArea>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
  ↪LanguageTechnology</ms:LTClassRecommended>
</ms:LTArea>
<ms:LTArea>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
  ↪MachineTranslation</ms:LTClassRecommended>
</ms:LTArea>
```

12.2.10 serviceOffered (M)

Path MetadataRecord.DescribedEntity.Organization.serviceOffered

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Lists the service(s) offered by an organization or person

Example

```
<ms:serviceOffered xml:lang="en">Evaluation and benchmarking</ms:serviceOffered>
<ms:serviceOffered xml:lang="en">Legal support</ms:serviceOffered>
```

12.2.11 domain (M)

Path MetadataRecord.DescribedEntity.Organization.domain

Data type component

Optionality Recommended

Explanation & Instructions

Identifies a domain that the organization deals with

You must fill in the CategoryLabel element with a free text value. If you prefer to add a value from an established controlled vocabulary, you can also use the DomainIdentifier (with the attribute DomainClassificationScheme with the appropriate value).

Example

```
<ms:domain>
  <ms:categoryLabel xml:lang="en">environment</ms:categoryLabel>
</ms:domain>
```

12.2.12 keyword (M)

Path MetadataRecord.DescribedEntity.Organization.keyword

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces a word or phrase considered important for the description of the project and thus used to index or classify it

Example

```
<ms:keyword xml:lang="en">Computational Linguistics</ms:keyword>
<ms:keyword xml:lang="en">Natural Language Processing</ms:keyword>
<ms:keyword xml:lang="en">Language Resources</ms:keyword>
<ms:keyword xml:lang="en">Research infrastructures</ms:keyword>
<ms:keyword xml:lang="en">Language Resources</ms:keyword>
<ms:keyword xml:lang="en">Digital Humanities</ms:keyword>

<ms:keyword xml:lang="en">Language Resources and Evaluation</ms:keyword>
<ms:keyword xml:lang="en">Legal support</ms:keyword>
<ms:keyword xml:lang="en">Data management</ms:keyword>
```

12.2.13 email (M)

Path MetadataRecord.DescribedEntity.Organization.email

Data type string

Optionality Recommended

Explanation & Instructions

Points to the email address of a person, organization or group

Example

```
<ms:email>info@elda.org</ms:email>
```

12.2.14 website (M)

Path MetadataRecord.DescribedEntity.Organization.website

Data type URL

Optionality Recommended

Explanation & Instructions

Links to a URL that acts as the primary page (like a table of contents) introducing information about an organization (e.g., products, contact information, etc.) or project

Example

```
<ms:website>https://www.cuni.cz</ms:website>
<ms:website>http://www.elra.info/en/</ms:website>
```

12.2.15 socialMediaOccupationalAccount (M)

Path MetadataRecord.DescribedEntity.Organization.socialMediaOccupationalAccount

Data type multilingual string

Optionality Recommended

Explanation & Instructions

Introduces the social media or occupational account details of a person or organization

You must also use the attribute `socialMediaAccountType` to specify the type of social media account. See https://european-language-grid.readthedocs.io/en/release1.0.0/Documentation/ELG-SHARE_xsd_Attribute_ms_socialMediaOccupationalAccountType.html#socialMediaOccupationalAccountType for details.

Example

```
<ms:socialMediaOccupationalAccount ms:socialMediaOccupationalAccountType="http://w3id.org/meta-share/meta-share/facebook">https://www.facebook.com/UFALMFFUK</ms:socialMediaOccupationalAccount>
```

12.2.16 hasDivision (M)

Path MetadataRecord.DescribedEntity.Organization.hasDivision

Data type component

Optionality Recommended

Explanation & Instructions

Links an organization to the division(s) it consists of

Example

```
<ms:hasDivision>
  <ms:divisionName xml:lang="en">Institute of Formal and Applied Linguistics</ms:divisionName>
  <ms:divisionShortName xml:lang="en">UFAL</ms:divisionShortName>
  <ms:divisionCategory>http://w3id.org/meta-share/meta-share/institute</ms:divisionCategory>
  <ms:organizationBio xml:lang="en">'Institute of Formal and Applied Linguistics (IFAL) at the Computer Science School, Faculty of Mathematics and Physics, Charles University, Czech Republic. The institute was established in 1990 as a continuation of the research and teaching activities carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later at the Faculty of Mathematics and Physics, Charles University. The Institute is a primarily research department working on many topics in the area of Computational Linguistics, and on many research projects both nationally and internationally. However, the Institute of Formal and Applied Linguistics is also a regular department in the sense that it carries a comprehensive teaching program both for the Master's degree (Mgr., or MSc.) as well as for a doctorate (PhD.) in Computational Linguistics. Both programs are taught in Czech and English. The Institute is also a member of the double-degree "Master's LCT programme" of the EU. Students also can take advantage of the Erasmus program for typically semester-long stays at other European Universities abroad.'</ms:organizationBio>
```


(continued from previous page)

```

    <ms:logo>https://ufal.mff.cuni.cz/sites/all/themes/drufal/css/logo/logo_ufal_
    ↪110u.png</ms:logo>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/
    ↪MachineTranslation</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/SpeechRecognition
    ↪</ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/Annotation</
    ↪ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassRecommended>http://w3id.org/meta-share/omtd-share/LexiconCreation</
    ↪ms:LTCClassRecommended>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Lexical Resources</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Dialog systems</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Corpus Creation</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Research Infrastructure</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>LT services</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>NLP Support</ms:LTCClassOther>
  </ms:LTArea>
  <ms:LTArea>
    <ms:LTCClassOther>Digital Humanities</ms:LTCClassOther>
  </ms:LTArea>
  <ms:keyword xml:lang="en">Computational Linguistics</ms:keyword>
  <ms:addressSet>
    <ms:address xml:lang="en">Malostranská 25</ms:address>
    <ms:zipCode>11800</ms:zipCode>
    <ms:city xml:lang="en">Praha 1</ms:city>
    <ms:country>CZ</ms:country>
  </ms:addressSet>
</ms:hasDivision>

```

Internal LT Service API specification

Note: This specification details the API that LT tool containers need to implement in order to be runnable as functional services within the ELG infrastructure. This is distinct from (though closely related to) the public-facing service execution API that outside users use to send requests to ELG services - the *public APIs* are documented separately.

Contents

- *Internal LT Service API specification*
 - *Basic API pattern*
 - *Utility datatypes*
 - * *Status message*
 - * *Annotations*
 - *Request structure*
 - * *Text requests*
 - * *Structured text request*
 - * *Audio requests*
 - *Response structure*
 - * *Failure message*
 - * *Successful response message*
 - * *Annotations response*
 - * *Classification response*
 - * *Texts response*

* *Audio response*

- *Progress Reporting*
- *Appendix: Standard status message codes*

Where possible, this document SHOULD use the MUST/SHOULD/MAY terms from [RFC 2119](#) to indicate requirement levels.

13.1 Basic API pattern

In order to integrate an LT tool as a functional service in the ELG infrastructure, the tool MUST offer at least one endpoint that can accept HTTP (1.1 or 2 - preferably cleartext HTTP/2) POST requests conforming to the appropriate request schema, and return an appropriate response as `application/json`. This specification also details a response pattern based on Server-Sent Events (SSE, a protocol defined as [part of HTML5](#)) that long-running tools can use to report progress information - support for this mechanism is RECOMMENDED for all tools but not required.

Endpoints may be sent multiple parallel requests by the ELG platform, and there is no requirement that a service must respond to requests in any particular order - certain services may, for example, be more efficient if they can batch up several requests into one back end process (e.g. for GPU computing) and send the responses in one go. If a tool has limits on the number of concurrent requests a single instance can handle then this information should be supplied to the ELG platform administrators as part of the on-boarding process, so the platform can use this data to decide how to scale the pod replicas to match the level of load on the service at any given time.

Where a tool already has its own native HTTP API it may be more convenient for integrators to provide a separate *service adapter* image which can handle requests matching the ELG specification and transform them into calls on the tool's native API. The tool container and the adapter container will run within the same "pod" in Kubernetes and can access each other as `localhost`.

13.2 Utility datatypes

The following JSON structures are used in several places in this specification, they are documented here to avoid duplication.

13.2.1 Status message

Since the ELG is supposed to be a multilingual platform, error and other status messages are handled using an approach modelled on the `i18n` mechanism from the [Spring Framework](#) - the message is represented by a *code*, along with a template *text* with numbered placeholders that are zero-based indices into an array of *params* replacement values.

```
{
  "code": "elg.example.no.translation",
  "text": "Default text to use for the {0} if no {1} can be found",
  "params": ["message", "translation"],
  "detail": {
    // arbitrary further details that don't need translation,
    // such as a stack trace, service-native error code, etc.
  }
}
```

ELG provides a common library of fully-translated message codes for service developers to use, as detailed below - developers are free to use their own codes in their own namespaces (i.e. not prefixed `elg.`) on the understanding that

it is their responsibility to provide translations. A mechanism for developers to contribute their translated messages to the platform is under development but not yet generally available.

13.2.2 Annotations

Many of the request and response types need to represent *annotations* - pieces of metadata about specific parts of a text or audio data stream, rather than about the stream as a whole. For example, a named entity recogniser might want to state that characters 10 to 15 in the request text represent the name of a female person, or a speech recogniser might want to state that characters 75 to 80 in the transcription represent a word, and map to the time period 1.37 to 1.6 seconds in the source audio. Such structures are represented in a consistent way across all the ELG API messages:

```
"annotations":{
  "<annotation type>":[
    {
      "start":number,
      "end":number,
      "sourceStart":number,
      "sourceEnd":number,
      "features":{" /* arbitrary JSON */ }
    }
  ]
}
```

The `<annotation type>` is an arbitrary string representing the type of annotation, e.g. “Person” or “Word” in the examples above. For each type of annotation, the matching value is a JSON *array* of objects, each object representing one annotation of that type. Note that when generating these structures in your API responses the value here **MUST** be an array even if there is only one annotation of the relevant type - some JSON generation libraries “unwrap” singleton arrays by default. The properties of each annotation object are:

start and end The position of the annotation in the main data stream to which it refers - this is typically the content directly associated with this `annotations` structure (for example the text of a translation). When the stream is text these would be Unicode character offsets from the start of the text, for audio they would typically be time points in seconds, etc. Subtracting the start value from the end value should give the length of the annotated area - there are several equivalent ways to conceptualise this, for example with text you could consider the characters as numbered from zero with the start offset *inclusive* and the end offset *exclusive*, or you could consider the offsets to represent the positions *between* characters (so 0 is before the first character, 1 is between the first and second, etc.).

sourceStart and sourceEnd Where these annotations are relative to a data stream that has been generated from another “source” data stream (e.g. a translation of text in another language, or a transcription of audio), these properties can be optionally used to link to the positions in the source stream (e.g. to align words in the translation with words in the original).

features Arbitrary JSON representing other properties of the annotation, e.g. a “Person” annotation might have a feature for “gender”, a “Word” from a morphological analyser might have “root” and “suffix”, etc.

13.3 Request structure

There are two main types of endpoint currently supported for this specification, one for services whose input is structured or unstructured *text* and one for services whose input is *audio*.

13.3.1 Text requests

Services that take plain text (or something from which plain text can be extracted, e.g. HTML) as their input are expected to offer an endpoint that accepts POST requests with `Content-Type: application/json` that conforms to the following structure.

```
{
  "type": "text",
  "params": { ... }, /* optional */
  "content": "The text of the request",
  // mimeType optional - this is the default if omitted
  "mimeType": "text/plain",
  "features": { /* arbitrary JSON metadata about this content, optional */ },
  "annotations": { /* optional */
    "<annotation type>": [
      {
        "start": number,
        "end": number,
        "features": { /* arbitrary JSON */ }
      }
    ]
  }
}
```

We expect that across the ELG from amongst the large number of possible and supported document types, a set of a smaller number of document types will emerge as being preferred and well supported (for example, plain text, HTML, XML - we do not intend to support binary formats such as PDF or Word as “text” requests, but may introduce other formats to this specification at a later date).

The only part of this request that is guaranteed to be present is the `type` (which will always be “text”) and the `content`. So a minimal request would look like this:

```
{ "type": "text", "content": "This is an example request" }
```

The optional elements are:

mimeType the MIME type of the content, if it is not simply plain text

params vendor-specific parameters - it is up to the individual service implementor to decide how (or indeed whether) to interpret these

features metadata about the input *as a whole*

annotations *as described above* - the `start` and `end` are Unicode character offsets within the `content` and the `sourceStart` and `sourceEnd` are ignored.

Tools that are able to accept text requests are RECOMMENDED to also offer an endpoint that can accept just the plain text (or other types of) “content” posted directly, and treat that the same as they would a message with the “content” property equal to the post data, the “mimeType” taken from the request `Content-Type` header, and no features or annotations. The “params” should be populated from the URL query string parameters. This endpoint will not be called by the ELG platform internally but it will make the service easier to test outside of the ELG platform infrastructure, and for open-source tools it will allow users to easily download and run the tool locally in Docker on their own hardware.

13.3.2 Structured text request

This is very similar to the plain text request, but for services that require some structure to their input, for example a list of sentences for some MT services, a list of words for a service that re-segments a stream of ASR output into

a list of sentences, etc. Again, services that accept this kind of input should provide a POST endpoint that accepts Content-Type: application/json conforming to the following structure:

```
{
  "type": "structuredText",
  "params": { ... }, /* optional */
  "texts": [
    {
      "content": "The text of this node", /* either
      "texts": [/* same structure, recursive */], // or
      // mimeType optional - this is the default if omitted
      "mimeType": "text/plain",
      "features": { /* arbitrary JSON metadata about this node, optional */ },
      "annotations": { /* optional */
        "<annotation type>": [
          {
            "start": number,
            "end": number,
            "features": { /* arbitrary JSON */ }
          }
        ]
      }
    }
  ]
}
```

The type will always be “structuredText”, params (optional) allows for vendor-specific parameters whose interpretation is up to the individual service implementor, and texts will always be an array of at least one JSON object. The texts property forms a recursive tree-shaped data structure, each object will be either a *leaf node* containing a piece of content or a *branch node* containing another list of texts.

Leaf nodes have one required property content containing the text of this node, plus zero or more of the following optional properties:

mimeType the MIME type of the content, if it is not simply plain text

features metadata about this node as a whole

annotations *as described above* - the start and end are Unicode character offsets within the content and the sourceStart and sourceEnd are ignored.

Branch nodes have one required property texts containing an array of child nodes (which may in turn be branch or leaf nodes), plus zero or more of the following optional properties:

features metadata about this node as a whole

annotations *as described above* - the start and end are array offsets within the texts array (e.g. "start":0, "end":2 would refer to the first and second children - treat them as zero-based array indices where the start is *inclusive* and the end is *exclusive*) and the sourceStart and sourceEnd are ignored.

Here is the simplest possible example of a structured text request representing two sentences, each with several words, with no features and no annotations.

```
{
  "type": "structuredText",
  "texts": [
    {
      "texts": [
        { "content": "The" }, { "content": "European" }, { "content": "Language" }, { "content":
↪ "Grid" }
```

(continues on next page)

(continued from previous page)

```

    ]
  },
  {
    "texts": [
      { "content": "An" }, { "content": "API" }, { "content": "example" }
    ]
  }
]
}

```

13.3.3 Audio requests

Services that accept *audio* as input (e.g. speech recognition) are slightly more complex, given the input data cannot be easily encoded directly in JSON. Audio services must accept a POST of Content-Type: `multipart/form-data` with two parts, the first part named “request” will be `application/json` conforming to the following structure, and the second part named “content” will be `audio/x-wav` or `audio/mpeg` containing the actual audio data.

```

{
  "type": "audio",
  "params": { ... }, // optional
  "format": "string", // LINEAR16 for WAV or MP3 for MP3, other types are service_
  ↳ specific
  "sampleRate": number,
  "features": { /* arbitrary JSON metadata about this content, optional */ },
  "annotations": { /* optional */
    "<annotation type>": [
      {
        "start": number,
        "end": number,
        "features": { /* arbitrary JSON */ }
      }
    ]
  }
}

```

The ELG platform typically expects audio to be a single channel - this is not guaranteed, as it depends what the requesting user submits, and a service receiving multiple audio channels may handle this situation in any way it sees fit including processing only the first channel or mixing down the multi-channel stream to mono before processing.

As with text requests we expect that there will be a small number of standard audio formats that are well supported across services (e.g. 16kHz uncompressed WAV) but individual services may support other types. The format and sample rate parameters may be ignored if the audio is in a format with a self-describing file header (e.g. WAV) which specifies other values.

Optional properties of this request type are:

params vendor-specific parameters - it is up to the individual service implementor to decide how (or indeed whether) to interpret these

features metadata about the input *as a whole*

annotations *as described above* - the start and end are floating point timestamps in seconds from the start of the audio and the sourceStart and sourceEnd are ignored.

13.4 Response structure

Services are expected to return their responses as JSON as described in the rest of this document. The minimal requirement is for services to be able to respond with `Content-Type: application/json` containing a successful or failed response message, but long-running services may also choose to offer `Content-Type: text/event-stream` to be able to stream progress reports during processing of the request. This mechanism is described at the end of this document.

13.4.1 Failure message

If processing fails for any reason (whether due to bad input, overloading of the service, or internal errors during processing) then the service should return the following JSON structure to describe the failure.

```
{
  "failure":{
    "errors":[array of status messages]
  }
}
```

The `errors` property is an array of *il8n status messages* (JSON objects with properties “code”, “text” and “params”) as described above - standard message codes are given in the appendix to this document.

13.4.2 Successful response message

All the successful responses follow this basic format:

```
{
  "response":{
    "type":"Response type code",
    "warnings":[/* array of status messages, optional*/,
    // other properties type-specific
  ]
}
```

As with the request, the response `type` code will likely be constant for any given service. The exact format of rest of a successful response message depends on the type of the service.

The `warnings` list is a slot to report warning messages that did not cause processing to fail entirely but may need to be fed back to the user (e.g. if the process involves several independent steps and only some of the steps failed, or the input was too long and the service chose to truncate it rather than fail altogether). Again, the individual messages in this array are *il8n status messages* as described above.

13.4.3 Annotations response

This response is suitable for any service that returns standoff annotations that are anchored to locations in text (e.g. named entity recognition) or time points in an audio/video stream (in general: anything compatible with a 1-dimensional coordinate system that uses a single number).

```
{
  "response":{
    "type":"annotations",
    "warnings":[...], /* optional */
  }
}
```

(continues on next page)

(continued from previous page)

```

    "features":{...}, /* optional */
    "annotations":{
      "<annotation type>":[
        {
          "start":number,
          "end":number,
          "features":{ /* arbitrary JSON */ }
        }
      ]
    }
  }
}

```

features (optional) metadata about the input *as a whole*

annotations (required, but may be empty "annotations": {}) *as described above* - for plain text data start and end would be character offsets into the text (Unicode code points), for audio data they would be the time point within the audio in seconds. The sourceStart and sourceEnd are ignored since there are no separate “source” and “target” data streams in this situation.

13.4.4 Classification response

For document-level (or more generally whole-input-level) classification services, e.g. language identification

```

{
  "response":{
    "type":"classification",
    "warnings":[...], /* optional */
    "classes":[
      {
        "class":"string",
        "score":number /* optional */
      }
    ]
  }
}

```

We allow for zero or more classifications, each with an optional score. Services should return multiple classes in whatever order they feel is most useful (e.g. “most probable class” first), this order need not correspond to a monotonic ordering by score - we don’t assume scores are all mutually comparable - and the order will be preserved by any subsequent processing steps.

Classification tools that classify *segments* of the input rather than the whole input should use the annotations or texts response formats instead of this one.

13.4.5 Texts response

A response consisting of one or more *new* texts with optional annotations, for example multiple alternative possible translations from an MT service or transcriptions from an ASR service.

```

{
  "response":{
    "type":"texts",
    "warnings":[...], /* optional */

```

(continues on next page)

(continued from previous page)

```

"texts":[
  {
    "role":"string", /* optional */
    "content":"string of translated/transcribed text", // either
    "texts":[/* same structure, recursive */],          // or
    "score":number, /* optional */
    "features":{" /* arbitrary JSON, optional */ },
    "annotations":{" /* optional */
      "<annotation type>":[
        {
          "start":number,
          "end":number,
          "sourceStart":number, // optional
          "sourceEnd":number,   // optional
          "features":{" /* arbitrary JSON */ }
        }
      ]
    }
  }
]
}
}
}
}

```

As with the structured text request format above, this texts response structure is recursive, so it is possible for each object in the list to be a branch node containing a set of child texts or a leaf node containing a single string.

Leaf nodes have one required property `content`, plus zero or more of the following optional properties:

role the role of this node in the response, “alternative” if it represents one of a list of alternative translations/transcriptions, “segment” if it represents a segment of a longer text, or “paragraph”, “sentence”, “word” etc. for specific types of text segment.

score if this is one of a list of alternatives, each alternative may have a score representing the quality of the alternative

features metadata about this node as a whole

annotations *as described above* - the `start` and `end` are Unicode character offsets within the `content` and the `sourceStart` and `sourceEnd` are the offsets into the source data (the interpretation depends on the nature of the source data).

Branch nodes have one required property `texts` containing an array of child nodes (which may in turn be branch or leaf nodes), plus zero or more of the following optional properties:

role the role of this node in the response, “alternative” if it represents one of a list of alternative translations/transcriptions, “segment” if it represents a segment of a longer text, or “paragraph”, “sentence”, “word” etc. for specific types of text segment.

features metadata about this node as a whole

annotations *as described above* - the `start` and `end` are array offsets within the `texts` array (e.g. “start”:0, “end”:2 would refer to the first and second children - treat them as zero-based array indices where the start is *inclusive* and the end is *exclusive*) and the `sourceStart` and `sourceEnd` are the offsets into the source data (the interpretation depends on the nature of the source data).

The texts response type will typically be used in two different ways, either

- the top-level list of texts is interpreted as a set of *alternatives* for the whole result - in this case we would expect the `content` property to be populated but not the `texts` one, and a “role” value of “alternative” - tools should return the alternatives in whatever order they feel is most useful, typically descending order of likelihood (though

as for classification results we don't assume scores are mutually comparable and the order of alternatives in the array need not correspond to a monotonic ordering by score).

- the top-level list of texts is interpreted as a set of *segments* of the result, where each segment can have N-best alternatives (e.g. a list of sentences, with N possible translations for each sentence). In this case we would expect `texts` to be populated but not `content`, and a “role” value of either “segment” or something more detailed indicating the nature of the segmentation such as “sentence”, “paragraph”, “turn” (for speaker detection), etc. - in this case the order of the texts should correspond to the order of the segments in the result.

13.4.6 Audio response

A response consisting of a piece of audio (e.g. an audio rendering of text in a text-to-speech tool), optionally with annotations linked to either or both of the source and target data.

```
{
  "response": {
    "type": "audio",
    "warnings": [...], /* optional */
    "content": "base64 encoded audio for shorter snippets",
    "format": "string",
    "features": { /* arbitrary JSON, optional */ },
    "annotations": {
      "<annotation type>": [
        {
          "start": number,
          "end": number,
          "sourceStart": number, // optional
          "sourceEnd": number,  // optional
          "features": { /* arbitrary JSON */ }
        }
      ]
    }
  }
}
```

Here the `content` property contains base64-encoded audio data, and the `format` specifies the audio format used - in this version of the ELG platform the supported formats are `LINEAR16` (uncompressed WAV) or `MP3`. In addition the response may contain zero or more of the following optional properties:

features metadata about this node as a whole

annotations *as described above* - the `start` and `end` are time offsets within the audio `content` expressed as floating point numbers of seconds, and the `sourceStart` and `sourceEnd` are the offsets into the source data (the interpretation depends on the nature of the source data).

As an alternative to embedding the audio data in base64 encoding within the JSON payload, a service MAY simply return the audio data directly with the appropriate `Content-Type` (`audio/x-wav` or `audio/mpeg`), however this approach means the service will be unable to return features or annotations over the audio, and will be unable to report partial progress.

13.5 Progress Reporting

Some LT services can take a long time to process each request, and in these cases it may be useful to be able to send intermediate progress reports back to the caller. This serves both to reassure the caller that processing has not silently failed, and also to ensure the HTTP connection is kept alive. The mechanism for this in ELG leverages the standard

“Server-Sent Events” (SSE) protocol format - *if* the client sends an `Accept` header that announces that it is able to understand the `text/event-stream` response type, then the service may choose to *immediately* return a 200 “OK” response with `Content-Type: text/event-stream` and hold the connection open (using chunked transfer encoding in HTTP/1.1 or simply not sending a `Content-Length` in HTTP2). It may then dispatch zero or more SSE “events” with JSON data in the following structure:

```
{
  "progress":{
    "percent"://number between 0.0 and 100.0,
    "message":{
      // optional status message, with code, text and params as above
    }
  }
}
```

followed by *exactly one* successful or failed response in the usual format. Services should not send any further progress messages once the success or failure response has been sent. Note that if a message is provided in a progress report it must be an *18n status message*, not simply a plain string.

For example:

```
Content-Type: text/event-stream

data:{"progress":{"percent":0.0}}

data:{"progress":{"percent":20.0}}

data:{"progress":{
data:    "percent":70.0
data:  }
data:}

data:{"response":{...}}
```

As per the SSE specification, *each line* of data within an event is prefixed `data:`, and an event is terminated by a blank line - there **MUST** be two consecutive newlines or CRLF sequences between the end of one event and the start of the next.

One would normally expect the progress percentage to increase over time but this is not necessarily a requirement of the specification - services are free to publish progress messages *without* a “percent” property if they wish to provide a status update message but cannot quantify their progress numerically, or even with a lower percentage than the previous message if they now have information to suggest that the overall process will take longer than first estimated.

Services are **RECOMMENDED** to support this response format, and to send it if the client indicates they can accept `text/event-stream`, but it is not required. The clients which will call your services within the ELG infrastructure will accept both `text/event-stream` and `application/json` responses, and you are encouraged to return an event stream if you can, but you are free to return `application/json` if it makes more sense for your service, and you **MUST** return `application/json` if the calling client does not indicate in the `Accept` header that they can understand `text/event-stream`.

13.6 Appendix: Standard status message codes

```
#
# Copyright 2019 The European Language Grid
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
# http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# This file contains the standard ELG status messages, translations should
# be placed in files named elg-messages_LANG.properties alongside this file.
#

# general bad request errors
elg.request.invalid=Invalid request message
elg.request.missing=No request provided in message
elg.request.type.unsupported=Request type {0} not supported by this service
elg.request.property.unsupported=Unsupported property {0} in request

elg.request.too.large=Request size too large

# Errors specific to text requests
elg.request.text.mimeType.unsupported=MIME type {0} not supported by this service

# Errors specific to audio requests
elg.request.audio.format.unsupported=Audio format {0} not supported by this service
elg.request.audio.sampleRate.unsupported=Audio sample rate {0} not supported by this_
↳service

# Errors specific to structured text requests
elg.request.structuredText.property.unsupported=Unsupported property {0} in "texts"_
↳of structuredText request

# General bad response errors
elg.response.invalid=Invalid response message
elg.response.type.unsupported=Response type {0} not supported

# Unknown property in response
elg.response.property.unsupported=Unsupported property {0} in response
elg.response.texts.property.unsupported=Unsupported property {0} in "texts" of texts_
↳response
elg.response.classification.property.unsupported=Unsupported property {0} in "classes_
↳" of classification response

# User requested a service that does not exist
elg.service.not.found=Service {0} not found

# generic internal error when there's no more specific option
elg.service.internalError=Internal error during processing: {0}
```

Public LT API specification

LT services can be called via the API endpoints given on the “code samples” page, the format of the various requests and responses is closely related to the internal LT service API used within the ELG infrastructure, but the public endpoints also offer shortcuts to simplify common interactions. There are three different types of endpoint depending on the kind of data required by the service as input - *flat text*, *structured text* or *audio*.

Authentication to all endpoints is by the use of an OAuth2 Bearer Token, and a token suitable for test use can be copied from the “code samples” page. Future versions of this document will include details of how to obtain and renew access tokens programatically. The token is passed via the HTTP `Authorization` header in the usual way:
`Authorization: Bearer <tokenValue>`

14.1 Input formats

14.1.1 Services that process flat text

`https://{domain}/execution/processText/{ltServiceID}`

Services that process a single flat stream of text can be called via an endpoint of this form. Make an HTTP POST request to the endpoint with one of the following `Content-Type` headers:

application/json A JSON object as described in the “*text request*” section of the LT Service API specification. For example `{ "type": "text", "content": "The text to process", "params": { "genre": "news" } }`. The type *must* be the string “text”, the content is the text to be processed, and `params` are specific to the individual service - see the per-service documentation for details of any parameters the service accepts.

text/plain or text/html Just the text to be processed. In this case any URL query parameters added to the endpoint URL will be passed on to the service as `params`

14.1.2 Services that process “structured” text

`https://{domain}/execution/processStructured/{ltServiceID}`

Some services require text that has been pre-segmented in some way, for example split into tokens, sentences or paragraphs. For this endpoint the following Content-Type values are supported:

application/json A JSON object as described in the “*structured text request*” section of the LT Service API specification. For example `{"type": "structuredText", "texts": [{"content": "First sentence."}, {"content": "Second sentence."}]}`. As with text requests above, you may also add `params` to the JSON, these are specific to the individual service - see the per-service documentation for details of any parameters the service accepts.

text/plain As a convenience shortcut the endpoint can also accept a POST of plain text. In this case, how the text is segmented depends on a URL query parameter `split`

processStructured/{service} (without a split parameter) the whole text is treated as a single segment

processStructured/{service}?split=line the text is divided at line breaks, and each line is treated as a separate segment. Leading or trailing white space on each line is *not* trimmed, and blank lines become empty segments `{"content": ""}`

processStructured/{service}?split=paragraph the text is divided at each run of one or more *blank lines* (i.e. two or more consecutive line breaks, possibly with white space in between). Again, leading or trailing whitespace around each segment is *not* trimmed.

All query parameters (including `split`) are passed on to the underlying service.

14.1.3 Services that process audio

`https://{domain}/execution/processAudio/{ltServiceID}`

Services that process a stream of audio can be called via an endpoint of this form. Make an HTTP POST request whose body is the audio data, with an appropriate Content-Type: `audio/mpeg` for MP3 audio or Content-Type: `audio/x-wav` for uncompressed WAV audio.

Any URL query parameters added to the endpoint will be passed on to the service, see the per-service documentation for details of which (if any) parameters the service accepts.

14.2 Service responses

The response formats returned from service calls are identical to *their counterparts in the internal LT Service API* and will not be repeated here. However there is one shortcut for services such as text-to-speech that return audio data. Ordinarily these services return a response of Content-Type: `application/json` including the audio data encoded in base64, but if you supply a parameter `audioOnly` (in the `params` for a JSON request, or as a URL query parameter for an unwrapped text/HTML/audio request) with the value “true” or “yes”, then instead of receiving the full JSON response you will receive just the binary audio data with an appropriate Content-Type of `audio/mpeg` or `audio/x-wav`.

Failed responses return a special type of response as follows:

```
{
  "failure": {
    "errors": [array of status messages]
  }
}
```

The `errors` property is an array of *internationalization-compatible status message objects* - the ELG platform provides another endpoint `https://{domain}/i18n/resolve` to which you can POST a JSON array of these objects and receive an array of resolved message strings in response.

14.3 Asynchronous processing

Some services may take several seconds or more to respond, either because their processing is naturally complex or because there are many requests for the same service being processed at the same time. To avoid the risk of dropped connections in such cases, the ELG platform offers an alternative “asynchronous” interaction style. To use this, send the same `POST` request, but add `/async` to the endpoint URL ahead of the `/process*`, e.g.

`https://{domain}/execution/async/processAudio/{ltServiceID}`

When called in `async` mode, the initial request should return immediately with a response of the following form:

```
{
  "response": {
    "type": "stored",
    "uri": "<polling URL>"
  }
}
```

The `uri` property is a URL which you should then begin to poll on a regular basis with a `GET` request (using the same `Authorization` token). Each time you poll, if processing is still ongoing you will receive a “progress” response of the form

```
{
  "progress": {
    "percent": //number between 0.0 and 100.0,
    "message": {
      // optional status message
    }
  }
}
```

(The `message` is optional, if provided it is a message *object* as in the failure response case above, which can be resolved to a message *string* by the `/il8n/resolve` endpoint). Some services return true progress percentages, for those that do not provide real updates the endpoint will always return `{"progress": {"percent": 0.0}}` to show that processing is still ongoing.

Once the processing is complete the poll URL will return the JSON response (successful or failed) exactly as you would have got from the normal synchronous API endpoint.

15.1 Build/Store Docker images

Ideally, the source code of your LT tool/service already resides on [GitLab](#) where a built-in [Continuous Integration \(CI\) Runner](#) can take care of building the image; GitLab also offers a [container registry](#) that can be used for storing the built image. For this, you need to add on the root level of your GitLab repository a `.gitlab-ci.yml` file as well as a `Dockerfile`; i.e., the recipe for building the image. [Here](#) you can find an example. After each new commit, the CI Runner is automatically triggered and runs the CI pipeline that is defined in `.gitlab-ci.yml`. You can see the progress of the pipeline on the respective page in GitLab UI (“CI / CD -> Jobs”); also when it completes successfully, you can find the image at “Packages -> Container Registry”.

Your image can also be built and tagged in your machine by running the `docker build` command. Then it can be uploaded (with `docker push`) to GitLab registry, in [DockerHub](#) (which is a public Docker registry) or any other Docker registry.

E.g for [this](#) GitLab hosted project the commands would be:

```
docker login registry.gitlab.com
```

For logging in and be allowed to push an image.

```
docker build -t registry.gitlab.com/european-language-grid/dfki/elg-jtok
```

For building an image (locally) for the project. Before running `docker build` you have to download (clone) a copy of the project and be in the top-level directory (`elg-jtok`).

```
docker push registry.gitlab.com/european-language-grid/dfki/elg-jtok
```

For pushing the image to GitLab.

In the following links you can find some more information on docker commands plus some examples:

- [Docker Command Line Interface](#).
- [Docker Tutorial from Stackify](#).

15.2 Dockerization of a Python-based LT service/tool

15.2.1 From Python script to Service

First you need to ensure that your python script provides either a std i/o messaging or a RESTful API.

A) create a REST API using **Flask** Example of a **Shakespeare Bot**

```
from flask import Flask
from flask import render_template
from flask import request

# creates a Flask application, named app
app = Flask(__name__)

from chatterbot import ChatBot

# a route where we will display a welcome message via an HTML template
@app.route("/", methods=['POST', 'GET'])
def hello():
    if request.method == 'POST': # this block is only entered when the form is_
        submitted
        user_message = request.form['user_message']
        chatbot = ChatBot("Frank")
        bot_message = chatbot.get_response(user_message).text

        data = {
            'bot_message': bot_message,
            'user_message': user_message,
            'user_message_visibility': '',
        }
        return render_template('index.html', **data)

    data = {
        'bot_message': "Speak. I am bound to hear.",
        'user_message': '',
        'user_message_visibility': 'style=visibility:hidden;',
    }
    return render_template('index.html', **data)

# run the application
if __name__ == "__main__":
    app.run(debug=True)
```

B) create a std i/o messaging Example to be added soon.

15.2.2 From Python Service to Docker image

Then you can make your service a Docker image by taking the following steps:

- choose python environment, e.g. python:3.6.4-slim-jessie
- add/copy python scripts
- add/copy other resources
- install missing modules

- define entrypoint

We provide you with some dockerfile examples to see its simplicity:

Example 1: Shakespeare Bot

```
from python:3.6.4-slim-jessie

COPY shakespearebot.py .
COPY corpora/hamlet.csv corpora/

RUN pip install pandas
RUN pip install chatterbot
RUN pip install chatterbot-corpus

ENTRYPOINT ["python", "shakespearebot.py"]
```

Example 2: Legal Entity Recognition, install all requirements from .txt-file

```
FROM python:3.7

COPY . .

RUN pip install -r requirements.txt

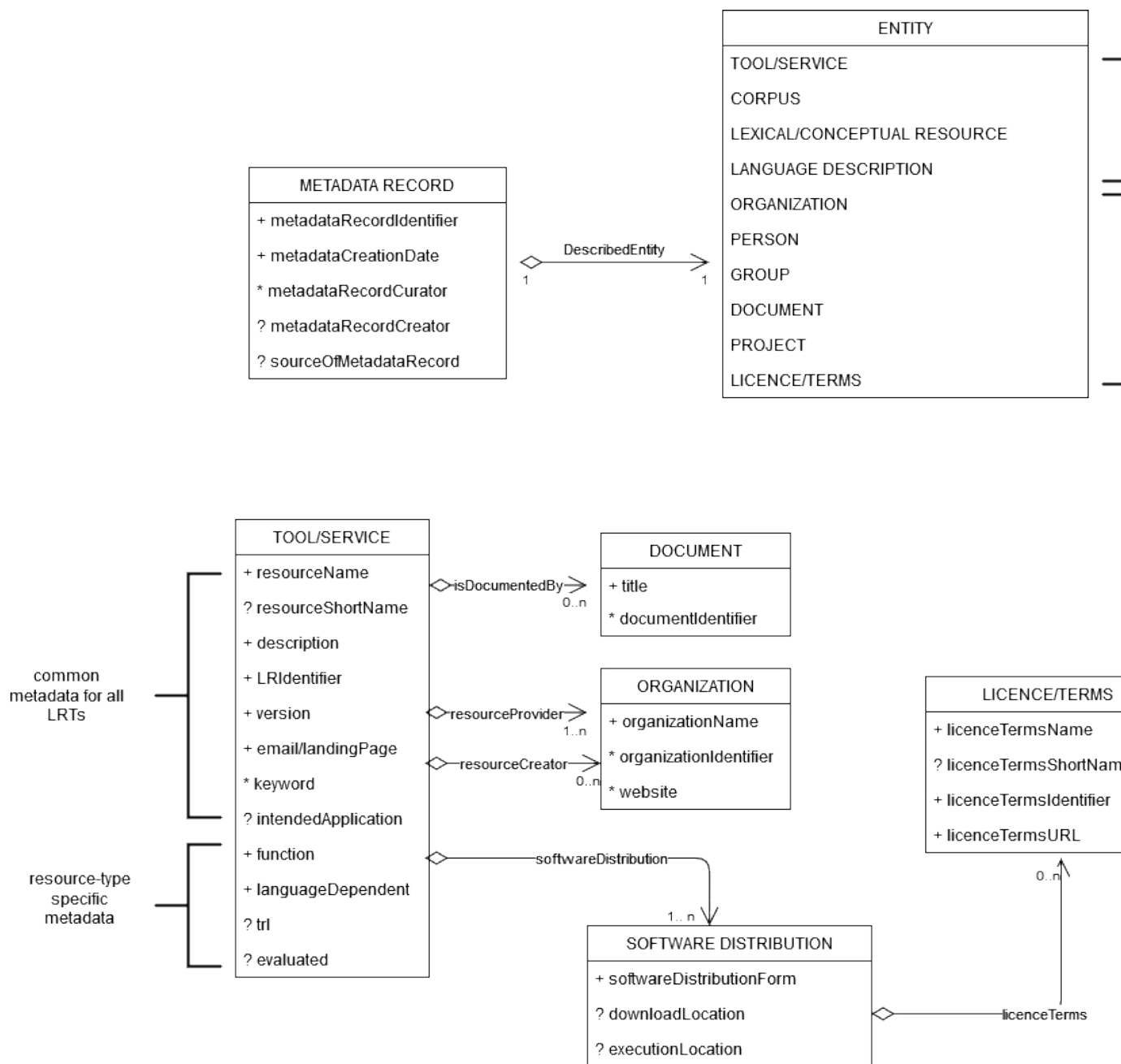
EXPOSE 8080
ENTRYPOINT ["python", "ler-ws.py"]
```

15.3 Dockerization of a Java-based tool

ELG Spring Boot Starter

16.1 Basic notions

The following figure shows the main notions upon which the ELG schema builds.



These include:

- **MetadataRecord**: It corresponds to the catalogue entry, and records information concerning the registration process, such as who created the entry and when, whether it was harvested from another catalogue, who is responsible for its curation (updates), etc.
- **DescribedEntity**: It corresponds to any entity that can be described by a metadata record. It can be a Language Resource, a Person, Organization, etc. (cf. [Overview](#)).
- **LanguageResource**, which is further classified into one of four resource types: **ToolService**, **Corpus**, **LexicalConceptualResource** and **LanguageDescription**. A Language Resource can be described through a set of metadata elements common to all, and a further set that fits to each of these four types.
- **Distribution**: It corresponds to the physical form with which a Language Resource is made available

through the catalogue, e.g. as a downloadable file, or a form accessed via an interface, etc.

16.2 Full schema documentation

You can find the full schema XSD, documentation as well as templates and examples of metadata records for all resource types in the [ELG SHARE schema Git repository](#).

You can browse the documentation here:

- [Metadata record \(Base item\)](#)
- [Language Resource](#)
 - [Tool/Service](#)
 - [Corpus](#)
 - [Language description](#)
 - [Lexical/Conceptual resource](#)
- [Project](#)
- [Organization](#)
- [Group](#)
- [Person](#)
- [Licence/Terms of use](#)
- [Document](#)

16.3 Minimal version

The minimal version comprises a set of carefully selected metadata elements that are deemed important for various reasons, such as:

- *identification and citation*: resource name(s); identifier(s); a short description of contents; versioning information; a contact point for further information (email or landing page); data of the resource provider(s) and resource creator(s); classification by domain, keywords and intended LT application; language coverage (language and, if needed, dialect); publication date;
- *support*: links to manuals, training material; samples of the resource;
- *usage/access*: distribution form (e.g. as downloadable file, a form that can be accessed via an interface, source code or binary file of software, etc.); licensing conditions; access location.

These metadata elements can be used to describe all resources, irrespective of the resource type. Additional metadata elements, particular to each resource type, are required, such as size and format for data files, prerequisites for tools and services, etc.

16.4 Template - Explanations

For each metadata element we present the following information:

- *Path*: the path of the element as in the XSD

- *Data type:*
 - string
 - multilingual string: you can repeat the element for different language versions; to specify the language, you must use the xml attribute `lang` with a value from IETF BCP 47, the [IANA Language Subtag Registry](#); for all metadata elements, a value in English (“en”) is mandatory
 - component: group of elements
 - Controlled Vocabulary (CV): value taken from a controlled vocabulary; a link to the relevant controlled vocabulary is provided
 - date: date in the format `xs:date`
 - URL
- *Optionality:*
 - **Mandatory (M)**: the element must always be filled in the metadata record
 - **Recommended (R)**: the use of the element is not enforced but provides important information
 - **Mandatory if applicable (MA)**: the element must be filled in when specific conditions apply
 - **Recommended if applicable (RA)**: the use of the element is recommended when specific conditions apply
- *Explanation & Instructions*: A short definition of the element, followed by instructions on how it should be used in the specific context.
- *Example*: One or more examples for the element in XML format.

CHAPTER 17

Indices and tables

- `genindex`
- `modindex`
- `search`

C

Code sample, [19](#)

D

Download, [8](#)

I

Internal LT API, [141](#)

P

Public API, [154](#)

T

Try Out, [19](#)