EUROPEAN LANGUAGE GRID

D2.3

Metadata schema

Authors:	Penny Labropoulou (ILSP), Katerina Gkirtzou (ILSP), Miltos Deligiannis (ILSP), Dimitris Galanis (ILSP), Maria Gavriilidou (ILSP), Stelios Piperidis (ILSP), Georg Rehm (DFKI), Maria Moritz (DFKI), Andrés Garcia Silva (EXPSYS)

Dissemination Level:	Public
Date:	31-08-2019

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 36 months
Deliverable number	D2.3
Deliverable title	Metadata schema
Type	Report
Number of pages	29
Status and version	Final – Version 1.0
Dissemination level	Public
Date of delivery	Contractual: 31-08-2019 – Actual: 31-08-2019
WP number and title	WP2: Grid Platform – Language Grid
Task number and title	Task 2.2: Metadata schema for the ELG Platform Catalogue
Authors	Penny Labropoulou (ILSP), Katerina Gkirtzou (ILSP), Miltos Deligiannis (ILSP), Dimitris Galanis (ILSP), Maria Gavriilidou (ILSP), Stelios Piperidis (ILSP), Georg Rehm (DFKI), Maria Moritz (DFKI), Andrés Garcia Silva (EXPSYS)
Reviewers	Ulrich Germann (UEDIN), Kalina Bontcheva (USFD)
Consortium	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Institute for Language and Speech Processing (ILSP), Greece
	University of Sheffield (USFD), United Kingdom
	Charles University (CUNI), Czech Republic
	Evaluations and Language Resources Distribution Agency (ELDA), France
	Tilde SIA (TILDE), Latvia
	Sail Labs Technology GmbH (SAIL), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Alexandru Ceausu
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Dr. Georg Rehm, DFKI GmbH georg.rehm@dfki.de Phone: +49 (0)30 23895-1833 Fax: +49 (0)30 23895-1810 http://european-language-grid.eu © 2019 ELG Consortium

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

API	Application Programming Interface
ASR	Automatic Speech Recognition
CEF	Connecting Europe Facility
CES	Corpus Encoding Standard
CMDI	Component MetaData Infrastructure
CMS	Content Management System
DoA	Description of Action
DMP	Data Management Plan
ELG	European Language Grid
ELRA	European Language Resource Association
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HTTP	Hypertext Transfer Protocol
IE	Information Extraction
IRI	Internationalized Resource Identifier
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation – Linked Data

LDC	Linguistic Data Consortium
LR	Language Resource
LRT	Language Resource and Technology
LT	Language Technology
MT	Machine Translation
MVP	Minimum Viable Product
NCC	National Competence Centre
NLP	Natural Language Processing
OLAC	Open Language Archives Community
OWL	Ontology Web Language
RDA	Research Data Alliance
RDF	Resource Description Framework
REST	Representational State Transfer
SME	Small and Medium Enterprise
TEI	Text Encoding Initiative
UI	User Interface
UML	Unified Modeling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VLO	Virtual Language Observatory
XML	Extensible Markup Language
XSD	XML Schema Definition

List of Terms

Functional content	Language processing tools and services that can be executed either locally or in a cloud infrastructure.
Language Data Resource, Language Data	A Language Resource composed of data, as opposed to a Language Technology tool or service.
Language Resource	A resource composed of linguistic material used in the construction, improvement or evaluation of language processing applications, but also, in a broader sense, in language and language-mediated research studies and applications; examples include datasets of various types, such as textual, multimodal or multimedia corpora, lexical data, grammars, language models, etc. in machine readable form. The term is often used in the bibliography and related initiatives with a broader meaning, encompassing also the (a) tools and services used for the processing

	and management of datasets, and (b) standards, guidelines and similar documents that support the research, development and evaluation of LT (cf. http://languagelog.ldc.upenn.edu/myl/ldc/LR_background.html). In this document, we use the term as first defined in the META-SHARE metadata model, i.e., including both data resources and Language Technology tools/services to avoid confusion with previous metadata descriptions. The alternative term "Language Resource/Technology" is also used.
Language Technology tool or service	A tool, service, or piece of software that performs language processing or any Language Technology related operation.
Linked Data	Linked Data is a structured data which is interlinked with other data so that they become more useful through semantic queries [Wikipedia, https://en.wikipedia.org/wiki/Linked_data]. It is based on a set of design principles for publishing and sharing machine-readable interlinked data on the Web, using URIs and RDF.
Metadata element, feature, field	Metadata is often defined as "data about data", i.e., data providing information about other data. A metadata element, feature or field is any piece of such information; e.g., title, author, date of creation, file format, etc.
Metadata schema, scheme, model	Metadata schemas define the structure of metadata elements, i.e., how metadata records can be set up. Datatypes, obligatoriness and cardinality of elements can also be part of a schema, among others.
Non-functional content	Any non-executable type of resource (data resource or source code of tools/services) that can be included in the European Language Grid.

Abstract

This document provides an overview of the ELG-SHARE Metadata Schema, which is used for the description of all entities included in the ELG catalogue, i.e., Language Technology Resources, both functional (tools and cloud-based services) and non-functional (corpora, lexica, terminologies, models, etc.), as well as entities related to them and involved in LT at large, such as persons, organizations, projects, documents and licences. The schema is implemented in the form of an XSD schema with metadata elements and values linked to the META-SHARE and OMTD-SHARE ontologies; its full documentation is available through the ELG portal. Following the dynamic ELG platform development process and feedback from users, the schema will be continuously updated; this deliverable is, hence, considered a living document (until M30), and as such it will be updated to reflect the evolution of the schema. This document presents the general framework of the schema, its design and implementation principles, and the metadata schemas and vocabularies it builds upon, and provides an overview of the main metadata information it contains for the entities it purports to describe.

1 Introduction

The ELG aspires to become the primary platform for LT in Europe, a one-stop shop for LT research, development, evaluation and commercial deployment. It will provide a catalogue of and access to

- LT resources, both functional (tools and cloud-based services) and non-functional (corpora, lexica, terminologies, models, etc.); and
- LT-related (meta-)information, such as information about relevant research and commercial organizations, events, training resources, job offerings, etc., as well as information about the aforementioned LT resources.

The collected information is formally structured and harmonized using the **ELG-SHARE Metadata Model** (or alternatively, in short, "ELG Metadata Model" or simply "Model")[1]. On the basis of the Model, metadata appropriately describing all LT assets are indexed and inventorized to create the ELG platform catalogue which serves as the entry point through which users access and deploy these assets.

D2.3 is organised as follows: Section 2 reports on the background and the factors that have influenced the design of the Metadata Model. Subsection 2.1 summarizes the user requirements with respect to metadata descriptions, while Subsection 2.2 presents the role of the Metadata Model in the ELG system. Subsections 2.3 and 2.4 provide a short overview of previous and current metadata-related activities that have been taken into account in the design of the Model. Subsection 2.5 describes the methodology used and the current status. Section 3 presents the entities covered by the Model and describes its main contents. Section 4 is devoted to issues around the formal implementation of the Model and the management of metadata records in ELG. Section 5 concludes with the remaining work. Finally, the full schema is documented in Annex A, and examples of metadata records are provided in Annex B; Annex C lists the people that have contributed to the Model.

[1] The terms "metadata schema" and "metadata model" are used interchangeably throughout the document.

2 Background and methodology of work

2.1 User requirements

In this section, we briefly outline the users' requirements (fully described in D2.1 and D3.1) that are directly related to the design of the ELG-SHARE Metadata Model. In line with the agile approach adopted for the user requirements, the Metadata Model will be refined during the course of the project to reflect updates in these requirements.

D2.1 classifies users of the ELG platform as follows:

- **content providers** and **developers and integrators** are providers and consumers of LT resources as described above;
- **information providers** and **information seekers** are providers and consumers of LT-related (meta)-information as described above;
- **citizens**, i.e., individuals that wish to be informed about LT in general, and understand the scope of ELG in advancing LT;
- **ELG platform and content administrators**, i.e., the ELG technical partners that will administer and monitor the day-to-day operation and performance of the platform and its content.

The design of the ELG-SHARE Metadata Model is affected by the requirements posed by all user types, albeit at a different degree and scope. For instance, the requirements of **content providers** and **developers** and **integrators** are the ones directly influencing the elements and values comprising the Model, given the fact that its main focus is the description of Language Resources and Technologies (LRTs) and related entities (such as persons, organizations, projects, etc.). On the other hand, the demand to serve **citizens** imposes simplicity in the presentation (structure) and mode of expression (e.g., definitions, examples) of the metadata elements.

In general, content providers are interested in the way they can best promote their products, services as well as themselves and their activities. In addition, researchers and research labs from academia and industry may want to provide exact details of resource creation for the sake of experimental reproducibility. As far as LT services are concerned, content providers want to be able to provide their tools and services in multiple ways, such as in the form of containerized tools that run in the ELG platform, but also in a downloadable form from the ELG platform, as an API that lives on an infrastructure controlled by the provider himself, or via links to source code or software image repositories, such as GitHub, GitLab, DockerHub, etc. Similarly, for the data resources, content providers may want to be able to provide their content directly via the ELG platform, or via links to external repositories.

Developers and integrators are mostly interested in the aspects of findability and usability. For data resources, they often use as search criterion the languages of the content, the licence of the resource as well as domain classification, while for LT services, they prefer to use the availability of the source code, the natural language(s) covered by the service, the licensing and access conditions and, to a lesser degree, the programming language(s) API's provided. Another important demand on the part of LRT consumers is the provision of samples for data resources and demo versions for LT services.

These user requirements feed into specific metadata elements and whether they should be obligatory or optional, as well as into the way these will be exploited and displayed in the GUIs for the content provider (i.e., metadata editor form) and consumer (e.g., facets, landing page setup).

2.2 Objectives and principles

The ELG-SHARE Metadata Model is the backbone of the ELG Platform. It is used in the back-end for the design and implementation of the database and index, and in the front-end for the interface with all users.

The ELG catalogue provides **LR consumers** with all mechanisms and functionalities required to browse through or search its contents, and access the LT tools/services or data resources, depending on and fully respecting their terms of use. Search functions include simple keyword search as well as faceted search on the basis of selected elements of the Metadata Model. Based on ELG metadata, functionalities of the platform also provide the user with indications on which language tools/services can potentially be applied on which language data. To serve **providers of LRTs,** the ELG platform offers all mechanisms and functionalities to appropriately describe, identify (by using/aggregating existing persistent identifiers), ingest and store their assets. Multiple channels are foreseen for their provision and storage, including metadata editors, batch ELG compliant metadata import and harvesting on the basis of agreed harvesting protocols. A full account of the platform features is provided in Deliverable D2.2.

The Model also contributes to the ELG Data Management Plan (DMP) for the formal description of the LRs that will be ingested and hosted in the ELG platform. For more details, see Deliverable D5.4.

The main objectives of the Model (to be further refined during the project course in order to accommodate the evolving user requirements) include the following. The Model should

- cover needs of discoverability and findability of all LT assets targeted by ELG
- satisfy documentation needs for all of them at different levels of granularity, covering to the full extent LR properties throughout the whole lifecycle of their production and consumption, while allowing for a threshold of information deemed indispensable according to user requirements
- address (at the metadata level) interoperability requirements of resources belonging to the same types (e.g., corpora, lexica, tools/services) and media (e.g., text, audio, video), but coming from different sources with different descriptions, as well as between resources of different types and media (e.g., between datasets and services to be used for their processing)
- facilitate accessibility by human users and, where possible/required, machines (e.g., by documenting direct download/execution locations instead of landing pages)
- provide an overview of the LT landscape allowing users to navigate through applications, products, datasets, actors, projects, etc. linked through the LT activities they are engaged in
- act, where appropriate, as a bridge between information provided in the catalogue (i.e., information about LRTs and the LT landscape) and the information content of the ELG portal, which includes, for instance, training and dissemination material intended to promote awareness of LT to users (citizens, SMEs, industry, etc.) less knowledgeable about it, information about LT events, etc.

To comply with the above objectives, the main principles and strategies employed in the design of the ELG-SHARE schema consist of the following:

- adopt and adapt the FAIR principles[2] [Wilkinson et al. 2016], relevant recommendations and best practices to the needs of ELG
- re-use or map to existing widespread metadata schemas and authority vocabularies, especially those focusing on LRTs
- recommend and promote documentation standardization efforts and policies, by proposing, where needed, community relevant vocabularies and metadata elements, and supporting recent trends and developments in the area of (meta)data (e.g., persistent identifiers, data citation practices, etc.)
- normalize user input where appropriate but also allow for free user input, depending on the specificities of each metadata element, and in response to current practices and balancing between preferences of (meta)data providers (free input) and applications (controlled input)
- be flexible enough to support varying degrees of documentation completeness through a tiered mandatoriness system for the metadata elements
- organize the metadata elements into a semantically coherent structure, accommodating common vs. particular features of resource/media types and attaching properties to the appropriate metadata entity.

2.3 Background on metadata initiatives for LRTs

The design and construction of documentation schemas catering specifically to LRs has a long tradition. Relevant activities started in corpus linguistics, initially catering for text corpora, (e.g., Text Encoding Initiative, TEI[3], and Corpus Encoding Standard, CES[4]), soon followed with various schemas for the documentation, annotation and representation of speech corpora, audiovisual resources, sign language resources, lexical and terminological resources, multilingual datasets, translation memories, etc. Alongside them, LR distribution agencies, such as ELRA[5] and LDC[6], have created their own documentation schemas for their catalogues, aiming to facilitate LR discovery and providing customer-oriented information in order to help them in selecting the appropriate resource(s). The Semantic Web has also fostered various activities on metadata specifically for LRs, including linguistic ontologies (e.g., GOLD[7]) and metadata models (e.g., LiMe[8] for linguistic resources published as Linked Data).

---

[2] The FAIR principles were officially published In 2016 in a scientific article [Wilkinson et al. 2016] and aimed to improve data management, sharing and usage as a response to the increasing uptake of the data-driven approach in science and subsequent demand on data. The authors outline ten principles (guidelines) targeting Findability, Accessibility, Interoperability and Reuse of digital assets. The most important accomplishment of these principles is that they emphasize and support machine-actionability, i.e., the ability of computational systems to Find, Access, Interoperate and Reuse data with none or minimal human intervention, making it possible to deal with the explosion of data amounts. For more information, see https://www.force11.org/group/fairgroup/fairprinciples, https://www.go-fair.org/fair-principles/ and https://www.egi.eu/about/newsletters/what-is-fair/.

[3] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

[4] https://www.cs.vassar.edu/CES/

[5] http://catalogue.elra.info/en-us/

[6] https://catalog.ldc.upenn.edu

[7] http://linguistics-ontology.org

[8] http://art.uniroma2.it/lime/

Operating under the auspices of the International Organization for Standardization (ISO), the TC37 (Technical Committee 37 – Language and Terminology) SC4 (Subcommittee 4)[9] is dedicated to the topic of Language Resource Management, developing standards for the annotation and representation of LRs. From these, the "ISO 24622 – Component Metadata Infrastructure (CMDI)"[10] focuses on metadata, specifying a model based on the notion of *components*, groups of semantically coherent elements that can be combined together to form *profiles* [Broeder et al., 2008]; elements themselves can be linked to elements of ontologies and controlled vocabularies (through the use of identifiers, such as IRIs). Thus, the model enables the construction of profiles for specific language resource types (e.g., text or speech corpora, lexica, etc.) shared by different research groups (e.g., social sciences scholars, historians, literature researchers, etc.) on the basis of common components and elements. The CMD model was initiated and is currently exploited in the CLARIN Research Infrastructure[11], which offers access to digital LRs mainly for scholars in the Social Sciences and Humanities; its role is instrumental in the homogenization of the LR catalogues of the CLARIN federated centres as well as those of external sources (e.g., OLAC, Europeana, etc.) that populate the Virtual Language Observatory (VLO)[12].

The META-SHARE Metadata Model[13] [Gavrilidou et al. 2012] was developed following the CMD approach with a focus on LRs in the domain of Language Technology. It caters for the description of language data resources, including multimodal corpora, lexical/conceptual resources, and computational grammars, as well as language-processing technologies. It has been based on previous metadata initiatives, and incorporated feedback from a wide group of NLP and LT experts. It is in use in the META-SHARE catalogue of LRs[14], adopted as-is or with modifications by a large number of organizations for creating new CMD profiles, and is currently also used in the ELRA catalogue[15]. It has given rise to a set of "application profiles", i.e., schemas based on it, with modifications, restrictions, extensions and updates in order to fit to requirements set by specific applications: ELRC-SHARE[16] [Piperidis et al. 2018] was designed for public domain text resources, OMTD-SHARE[17] [Labropoulou et al. 2018] includes an extension to scholarly publications and Text and Data Mining workflows, while the version for the CEF-AT Catalogue of Services[18] is restricted to eTranslation services. The original META-SHARE schema has been converted into an OWL ontology[19] [McCrae et al. 2015] and as such used in LingHub[20] and the ReTeLe-SHARE ontology[21], a localized version in Spanish for the catalogue of Spanish LRs[22].

---

[9] https://www.iso.org/committee/297592/x/catalogue/

[10] https://www.iso.org/standard/37336.html?browse=tc

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

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

[13] http://metashare.ilsp.gr/knowledgebase/homePage

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

[15] http://catalogue.elra.info/en-us/

[16] https://gitlab.com/ilsp-nlpli-elrc-share/elrc-share-repository/tree/master/misc/schema/ELRC2

[17] https://github.com/openminted/omtd-share-schema, http://w3id.org/meta-share/omtd-share/ (partial implementation of the schema)

[18] https://cef-at-service-catalogue.eu

[19] http://purl.org/net/def/metashare

[20] http://linghub.org/

[21] https://w3id.org/def/retele-share

[22] http://catalogo.retele.linkeddata.es/

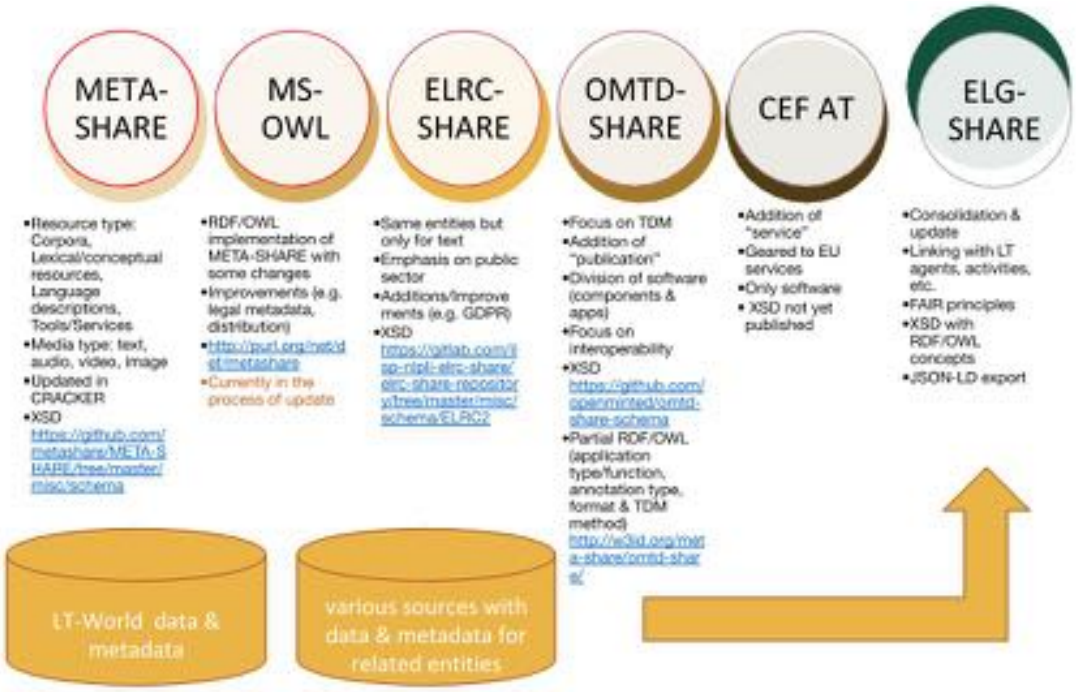

Figure 1: The main sources of the ELG-SHARE Model

For ELG, we decided to base the Metadata Model on two main source lines, as depicted in Figure 1:

- for the description of LRTs, we have selected META-SHARE and its application profiles, given their widespread use in the LT domain. The ELG-SHARE Model builds upon and extends them: modifications, updates and extensions in the contents (metadata elements and values) are made in response to user requirements (Section 2.1) and new descriptive needs (e.g., for GDPR-related features, billing module for commercial services, technological advances mainly in the description of containerized tools/services, improved description of ML models, etc.), taking also into account more recent developments in the metadata area at large (Section 2.4). In addition, a change in the implementation is made: as discussed in Section 4, previous META-SHARE profiles were implemented in the form of XSD (XML Schema Description) and imported/exported XML files; the ELG-SHARE combines the XSD approach with RDF (Resource Description Framework) specifications and introduces the JSON-LD serialization format for import/export purposes, moving closer to Linked Data and the Semantic Web technologies.
- for the enrichment of the descriptive model of the *satellite entities* that are related to LRs (e.g., actors responsible for the creation and curation of resources, funding projects, etc.), we build upon LT-World [Jörg et al. 2010], as well as other catalogues and datasets[23]. LT-World was an ontology-driven web portal aimed at serving the global LT community and providing information on actors, projects, events, resources, products, etc.; although the portal is no longer active, we plan to adapt the ontology and portal contents to the ELG objectives, especially for the purpose of bootstrapping the catalogue.

[23] For instance, FLaReNet reports on LR players (http://www.flarenet.eu), the LT-Innovate Directory of LT (http://www.lt-innovate.org/directory), relevant datasets from the EU Open Data portal (http://data.europa.eu/euodp/en/home), etc.

2.4 Overview of current (meta)data initiatives

There are various initiatives promoting best practices for the publication and sharing of digital data over the web including recommendations concerning metadata. The most notable ones among them are associated with the FAIR principles, the Open Science movement[24], Data Citation Principles[25], software citation principles[26], RDA recommendations and outputs[27] and OpenAIRE recommendations related to datasets and software[28]. ELG closely follows these recent developments and takes them into account for the design and implementation of the platform and, in so far as metadata issues are concerned, for the ELG-SHARE schema.

To illustrate the issues, we present here some of the requirements imposed by the FAIR principles on metadata and the way they are handled in the Model[29]:

- *F1 – (meta)data are assigned a globally unique and eternally persistent identifier*: the Model includes for each of the entities it targets (i.e., not only for LRs, but also for persons, organizations, etc.) an *identifier* element with an attribute for the *identification scheme* according to which they are assigned (i.e., the authority that has issued it); a separate *identifier* element is also included for the metadata record that describes them
- *F2 – data are described with rich metadata*: the Model includes elements for the whole lifecycle of the LRs, from production to consumption, the description of their contents and related entities; for the documentation of related entities, the set of elements is extended, covering all ELG user requirements
- *F4 – metadata specify the data identifier*: the *identifier* element for each of the entities described in ELG is included inside the metadata record
- *I1 – (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation* & *I2 – (meta)data use vocabularies that follow FAIR principles*: the Model makes use of the META-SHARE and OMTD-SHARE OWL ontologies
- *I3. (meta)data include qualified references to other (meta)data*: the Model includes a set of qualified (defined in the META-SHARE ontology) relations between the various entities that can be used when

[24] "Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods." [https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition]. See also https://ec.europa.eu/research/openscience/index.cfm and the European Open Science Cloud (https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud)

[25] As part of the movement for reproducible research and data reuse, data created and/or deployed in scientific workflows have been acknowledged as research assets of their own that must be cited, just like other sources and evidence. The Data Citation Principles are a set of guiding principles for citing data within scholarly literature, or any other dataset, or research object. For more information, see https://www.force11.org/datacitationprinciples.

[26] As a follow-up of the Data Citation Principles, the Force11 Software Citation Working Group has developed a set of principles for the citation of software; see https://www.force11.org/software-citation-principles

[27] The Research Data Alliance (RDA) is a community-driven initiative aiming to build the social and technical infrastructure to enable open sharing and re-use of data. The work is carried out by groups composed of data producers, users and stewards, and covers all data lifecycle stages, addressing data exchange, processing, and storage. The RDA endorsed outcomes can be found at https://www.rd-alliance.org/recommendations-and-outputs/all-recommendations-and-outputs

[28] OpenAIRE (https://www.openaire.eu) is an infrastructure dedicated to promoting and facilitating openness in scholarly literature and research, and contributing overall to the Open Science movement.

[29] Here we present only the FAIR principles that are directly relevant to the schema; principles about metadata that have an impact on technical issues (e.g., registration of metadata in a searchable resource) are outside the scope of this report.

describing a resource (e.g., establishing links between a raw corpus and its annotated version, a resource and its creator(s) or provider(s), etc.); as regards the metadata elements themselves, links to popular schemas, ontologies and vocabularies are also included in the ontology (see Section 3)

- *R1.1 – (meta)data are released with a clear and accessible data usage license*: the schema includes the *licence* entity which is separately described with a set of its own elements (at least a name, URL with the text, identifier(s), etc.); all other resources described by the schema include a *licence* element that links to it
- *R1.2 – (meta)data are associated with their provenance:* the schema includes a rich set of elements recording the provenance of the resource and the metadata (separately), providing information on the creation/modification/enrichment/processing procedures, tools/services used for them, dates when these happened, actor(s) involved in them, etc.
- *R1.3 – (meta)data meet domain-relevant community standards*: the schema is based on previous metadata schemas already well established in the LRT community and consolidated through the involvement of domain experts and the user community (cf. Section 4).

In a similar way, we take into account other recommendations regarding metadata and adapt them to the ELG requirements and specifications.

2.5 Working methodology and current status of the model

The ELG-SHARE model builds upon previous models (mainly META-SHARE and LT-World, see Section 2.4), as well as the ELG user requirements (cf. Section 2.1, D2.1 and D3.1). The reuse of pre-existing popular metadata models allows us to quickly have a fairly mature version of the model that can be used to build the platform database and catalogue for the initial platform release due in M16 (D2.4).

The current version of the schema (cf. Annex A) is the product of discussions carried out within the Metadata Task Force of the ELG Consortium (see Annex C). To facilitate the work, an initial set of UML-like diagrams[30] was created with all features required for the description of LRs and their related entities, based on the previous models and the user requirements. The outcomes of these discussions have led to an update of the META-SHARE[31] and OMTD-SHARE ontologies, which provide the building blocks (metadata elements and values) of the Model. Cardinality features and application-specific properties have been added in the XSD that implements the schema as described in Section 4.

Pre-releases of the two ontologies and the Model have been made available for review and feedback initially to a small group of experts from the Linked Data community involved in a collaborating project, Prêt-à-LLOD [32], since they had also worked on the META-SHARE ontology and were acquainted with its basic principles and features (see Annex A). The discussions will soon be moved to a wider circle, the Linked Data for Language

[30] The diagrams have been based on the UML entity-relationship diagram; some non-standard symbols and acronyms have been used, e.g., for defining datatypes for properties of the entities (e.g., "CV", standing for Controlled Vocabulary for enumerated lists, without further information), and colours to highlight specific properties that required discussion.

[31] To be published under the namespace http://w3id.org/meta-share/meta-share

[32] https://www.pret-a-llod.eu

Technology (LD4LT) W3C Community Group[33]. In addition, part of the OMTD-SHARE ontology, namely the taxonomy of Language Technologies (see Section 0), has also been presented at the LT-Summit[34]. There are ongoing discussions for forming an LT-Innovate Interest Group around this theme.

During the next few months, we intend to share the Model with a wider group of experts from the Grid community, mainly NCCs (National Competence Centers), and the ICT-29b projects, as well as other potential users, through face-to-face dissemination events and digital communication. This series of consultations aims to ensure a wide take-up of the model and give us feedback from prospective user communities.

According to the Description of Action, three platform releases are planned (cf. Table 1). Although the Model, in its current version, is meant to cover all features of the platform, there may be the need to add specific metadata elements/values or to make minor modifications. The updated versions of the Model, which will also consolidate the feedback we receive from users and above-mentioned community groups, will be delivered together with the respective platform releases.

Release	Features	Due
D2.4 ELG platform (first release)	• backend components required for the operation of the catalogue: simple user management component, components supporting documentation/uploading, storing/downloading of all resource types (tools and services, datasets, etc.), APIs required for interacting with other layers • first version of the guidelines on its use and provision of resources, instructions for containerization and invoking of remotely accessible web services • limited sets of tools and services and LRs	M16
D2.5 ELG platform (interim release)	• updated version of the platform including the components and APIs required for running language processing services (containerized services stored in the ELG and web services via REST APIs) directly in ELG • updated (as/if needed) version of the guidelines on its use and provision of resources, instructions for containerization and invoking of remotely accessible web services • updated catalogue with resources from ELG partners	M26
D2.6 ELG platform (final release)	• updated version of the platform including management and maintenance of the platform: monitoring of the platform, monitoring of remotely offered services, platform usage analytics, prototype version of user billing and payment services • updated catalogue with resources from ELG pilots and collaborating initiatives	M34

Table 1: ELG Platform release plan

[33] https://www.w3.org/community/ld4lt/

[34] "The Language Technology Market and Components Taxonomy", https://www.lt-summit.com

Finally, we should note here that in parallel with the ongoing discussions for the full Model, we have also implemented a limited version, restricted to a subset of the mandatory and recommended elements pertaining to tools/services and text corpora. This is used for the design of the database and a draft documentation of the entries that will populate the initial catalogue for the Minimum Viable Product (MVP) and was motivated by the need to produce in a short time a version of the ELG platform that can be used for demo and proof-of-concept purposes.

3 Presentation of the Metadata Model

3.1 Entities covered

ELG designs, develops and will deploy and populate the ELG platform *with* and *for* commercial and non-commercial Language Technologies alike, including both **functional** (running services and tools) and **non-functional** (corpora, lexica, terminologies, models, etc.) **resources**, as well as **LT-related meta-information**, e.g., information about research and commercial organizations providing LT solutions, LT initiatives and projects, business applications, etc. Table 2 presents in a concise way the different types of ELG offerings.

Type	Description	ELG will initially be populated with
Services and tools	Users of the ELG will be enabled to register, describe, upload, search and deploy as well as integrate containerized services, from simple tokenization or part-of-speech tagging to complex processing workflows.	GATE, GATECloud, UDPipe, TILDE's and University of Edinburgh's MT services, Expert System IE tools, SAIL LABS ASR, KWS, sentiment, age and gender detection tools, etc.
Remote services and tools	Users of the ELG will be enabled to register, describe, search and integrate functional remote APIs.	Research and commercial services, among others, GATECloud with 65+ services, UDPipe, TILDE's and University of Edinburgh's MT services, SAIL LABS and EXPERT SYSTEM services
Language resources, services, tools, software code	Users will be enabled to upload, describe, search, download datasets, source code packages, trained models, embeddings, etc.	Datasets, models, tools, lexica, etc. from META-SHARE, ELRC-SHARE, and ELRA catalogue, ELRA's services for e-licensing and e-Commerce
Information about services, tools, resources	Non-functional catalogue entries, for example, on a Language Technology provider company or research centre.	Information collected through projects and initiatives such as, among others, CRACKER, ELRA, ELRC, META-SHARE, LT World, etc.

Table 2: ELG layers of content

The ELG Metadata Model covers the description of the following entities (Figure 2):

- **Language Resources**, including data resources and language processing tools/services (functional as well as non-functional), which are the main focus of ELG, and
- **related/satellite entities**, that are involved in the lifecycle of LRs:

- **actors**, i.e., **persons, groups** or **organizations,** that have, e.g., created or curate the resources
- **documents,** such as publications that describe a resource or how it has been used for an application, user and training materials, etc.
- **projects**, that have, for instance, funded them or in which they have been used
- **licences/terms of use** that are used for the distribution of the resources.

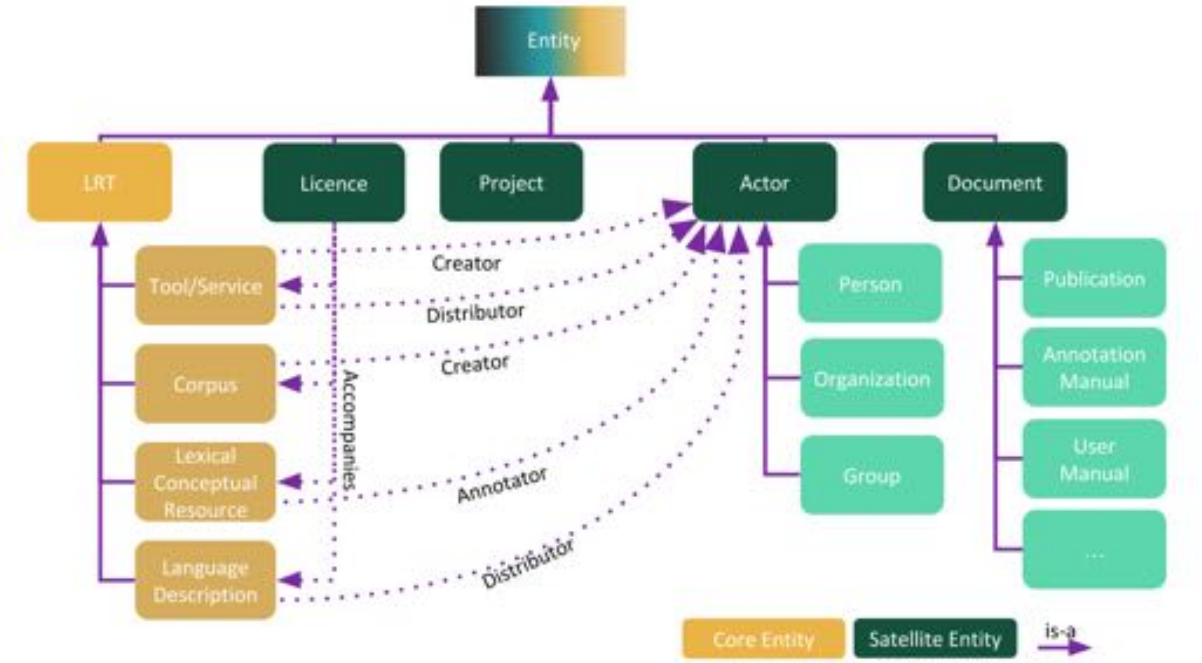

Figure 2: ELG-SHARE entities

3.2 Contents of the Model

The Model is composed of metadata elements that are used to describe *properties* of the entities described in the previous section and *relations* between them. Each of the entities can be described through its own set of metadata elements briefly outlined here and fully described in Annex A. In addition, all entities described in the ELG Model are also assigned a system-internal identifier[35] which is used for their interlinking.

When an entity appears for the first time as dependent on the description of another entity (e.g., a tool used for annotatating a resource, a person or organization that has created a resource, an author of a publication describing a resource, etc.), only minimal information is required for its encoding, namely the name or title and, optionally, an identifier and the identification scheme (e.g., ORCID, DOI, ISLRN, etc.) according to which it has been assigned. Metadata creators/curators can also enrich the description of these entities if they wish to by creating metadata records specifically for them.

[35] System identifiers are not meant to replace the persistent unique identifiers which are required for each entity according to FAIR principles and which are also covered by the model; their main aim is to ensure appropriate linking between entities, by correctly identifying and resolving them inside the ELG system.

The Model, as regards **related entities (i.e., actors, projects, etc.)** includes features required for their identification (i.e., at least name/title and identifier) and all necessary information describing relevant activities and products (e.g., links to LRTs, logos, promotional material, demos, etc.). Elements for additional information (e.g., contact person for an organization, link to a landing page for an organization or project, etc.) are also foreseen. Finally, where appropriate, GDPR issues for persons will be properly handled by the platform at the metadata creation steps (i.e., asking for permissions) and at display (with monitoring).

The ELG-SHARE Model, as regards **data resources and processing tools/services**, caters for their full life-cycle, integrating information about the resource identification, technical features, deployment requirements, legal rights and obligations, associated documentation, classification, etc.

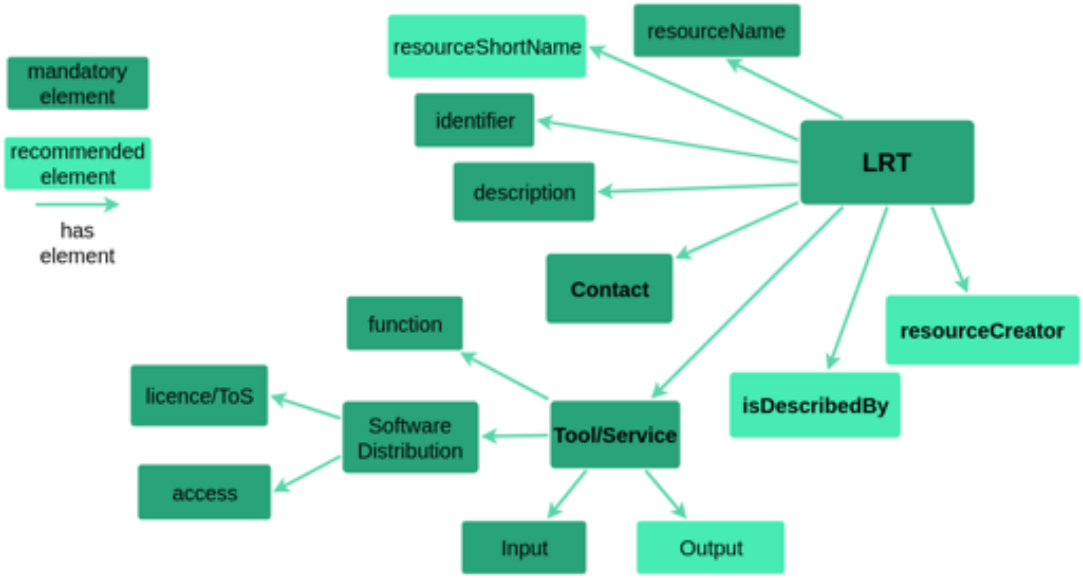

Figure 3: Simplified subset of the ELG Metadata Schema for LRTs

For the organization of the information of LRs, the model relies on three concepts (resource type, media type, distribution), each of which is associated with a distinctive set of metadata elements:

- **resource type**, with the following values:
  - **corpus**, defined for our purposes as a structured collection of pieces of data (textual, audio, video, multimodal/multimedia, etc.), typically of considerable size and selected according to criteria external to the data (e.g., size, type of language, type of text producers or expected audience, etc.) to represent as comprehensively as possible the object of study
  - **lexical or conceptual resource**, i.e., a resource (such as terminological glossary, word list, semantic lexicon, ontology, etc.) organized on the basis of lexical or conceptual units (lexical items, terms, concepts, phrases, etc.) with their supplementary information (e.g., grammatical, semantic, statistical information, etc.)
  - **language description**, i.e., a resource aiming to describe a language or some aspect(s) of a language via a systematic documentation of linguistic structures (e.g., computational grammars, statistical and machine learning-computed language models)
  - **tool or service**, covering all software that performs language processing and/or any Language Technology related operation (e.g., basic processing tools, applications, web services etc. that perform annotation, Machine Translation systems, speech recognizers, etc.)

- the **media type** element, which specifies the form/physical medium of the resource. The notion of medium is preferred over the written/spoken/multimodal distinction, as it has clearer semantics and allows us to view LRs as a set of modules, each of which can be described through a distinctive set of features. Thus, the following media type values are foreseen: **text, audio, image, video** and **numerical text** (referring to numerical data, such as biometrical, geospatial data, etc.)
- the **distribution**, which, following the DCAT[36] model, refers to any physical form of the resource that can be distributed and deployed by end-users.

Some elements, especially those that pertain to administrative features (e.g., identification, contact, funding information, etc.), are common to all types of resources, while other elements, mainly those representing technical features about the contents and format of resources, differ across resource and media types. For instance, a corpus consists of sets of files (e.g., whole texts) representative of a certain language, dialect or domain that can be processed with an annotation service, while a lexical or conceptual resource includes a set of units and their accompanying information (e.g., lemmas and their definitions or translation equivalents).

A resource may consist of parts belonging to different types of media: for instance, a multimodal corpus includes a video part (moving image), an audio part (dialogue) and a text part (subtitles and/or transcription of the dialogue); a multimedia lexicon includes the text part, but may also include a video and/or an audio part; a sign language resource is also a resource with various media types (video, image, text). Similarly, tools can be applied to resources of different media types, e.g., a tool can be used for processing both video and audio files.

Technical features, such as format, size or form of distribution, as well as licensing and billing information are properties of the distribution. And they can also differ across resource and media type. As an example, corpora can be distributed as PDF files or as simple text files, lexical resources in tabular forms or queried through an interface, while tools may be available as source code, as executable files or as web services. And each of these forms can be licensed under different terms: source code may be available at a price for integration in other applications, while an API may be offered for research purposes without any fee.

In the structure of the Model, metadata elements are attached to the appropriate level (element)[37]:

- common, mainly administrative and descriptive, features at the resource level: **identification** features (e.g., name of the resource, short description, identifiers, logo, etc.), information on the **version**, **contact** information (e.g., a landing page providing more information on the resource, or a contact person), information about the **creation** phase (e.g., funding project, resource creator), some **legal** information that relates to the resource as a whole (e.g., IPR holder, resource provider), features about the **validation** process (if any) of the resource, **classification** information that can be assigned to all resource types (keywords, domain), the **use** of the resource (e.g., project or application where it has

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

[37] In the ELG-SHARE schema, we have opted for an approach following the Linked Data paradigm instead of the Component Metadata approach, hanging *elements* (that represent *properties*) directly on the *elements* (that represent entities) without any semantic grouping – for further discussion, see Section 4.1.

been used, link to a publication that has used it), and **documentation** information (e.g., links to user manuals, publications describing the resource, training resources, demo videos, etc.)

- separate sets of features at the resource type level and, when required, separate resource/media type combinations (e.g., specific to corpus text parts vs. corpus audio parts or lexical/conceptual resource text parts), broadly covering (and not limited to) the following types of information:
    - **contents**: elements mainly referring to languages covered in the resource, types of content (e.g., for images: drawings, photos, histograms, animations etc.), modalities included (e.g., written or spoken language, gestures, eye movements, etc.), etc.
    - **classificatory** information, including resource-type subclassification (e.g., subtypes of lexical/conceptual resources, functions/tasks performed by tools/services etc.) as well as classification of the contents of the resource; this can be media-independent (e.g., geographic and time coverage) and media-dependent (e.g., text type, audio genre, setting of a video, etc.)
    - information related to **technical process of the creation** of specific resource parts, e.g., the original source, the capture and recording methods (e.g., scanning and web crawling for texts vs. recording methods for audio files)
    - **performance** of the resource, which is resource-type driven, given that the measures and criteria differ across resource types
    - **operational requirements** of the resource (e.g., the hardware and software prerequisites for running a tool/service, input and output of a tool/service providing information on the media type, format, language, etc. that the tool/service can take as input and the resulting output)
    - **relations** to other resources and entities, that mainly have to do with processing operations (e.g., relation between an annotated version of a corpus and the raw corpus, the annotation tool/service used, the annotator, etc.)
- sets of features for distributions of data resources and tools/services, including a common set of features regarding mainly **licensing terms** and related information (e.g., billing, distribution rights holders, copyright statement, etc.), and a separate set recording **technical descriptions** of the distributions: the distribution form, the location where a resource can be accessed/downloaded from, information on the (media type dependent) format of a resource, character encoding, size, etc.

When describing a resource in ELG, depending on the resource/media type combination, the appropriate set of features will be displayed to the metadata creator, e.g., for a spoken corpus and its transcriptions, the audio feature set will be used for the audio part and the text feature set for the transcribed part.

Due to the richness of information foreseen, the Model is complex. To ensure flexibility and uptake, the elements described in the next subsections are classified into three levels of optionality[38]:

- **mandatory:** elements that are necessary for the intended purposes in the ELG platform
- **recommended:** elements that can help the current or future use of the entity description, or useful information that has not yet been standardized

[38] Only mandatory vs. optional elements can be formally represented in XSD; recommended elements are marked as such in the documentation.

- **optional:** all remaining information.

This enables "relaxing" the metadata requirements by adhering to two versions of the schema:

- an initial level providing the basic elements for the description of a resource (**minimal schema**), including all mandatory elements and
- a second level with a higher degree of granularity (**maximal schema**), providing detailed information on an entity, including recommended and optional elements.

Finally, we should mention here one metadata element enjoying a special status in the Model, as it is used for interconnecting all entities in the ELG catalogue: the **Language Technology application area**. The element is organized as a *class* in the ontology[39] (Figure 4), i.e., with synonyms and hypernyms (hence the term "LT taxonomy" used for referring to it) and the possibility to add relations to similar external vocabularies and ontologies maintained by other communities.

Figure 4: Part of the LT taxonomy

[39] Given that work on the taxonomy had started for the OMTD-SHARE schema, with a focus on Text and Data Mining, we have decided to keep it under the original namespace (i.e., http://w3id.org/meta-share/omtd-share/), although it is currently extended to include all LTs envisaged in ELG.

The LT taxonomy can be regarded as a *controlled vocabulary* normalizing the metadata providers' input text and, thus, facilitating the indexing and discovery of resources. Through the exploitation of synonyms and hypernyms, the retrieval of results can also be improved. In addition, metadata providers will have the opportunity to add their own values in a free text element; these values will be considered as candidates for addition in the taxonomy, following a curation procedure that will be discussed during the project.

The platform will also support access to the catalogue contents through the LT taxonomy (i.e., as if the LT area was a proper entity like LRs, actors, etc.). This functionality accomplishes two objectives:

- raising awareness and promoting LT among the field experts, by providing an overview of the LT activities in relation to various criteria (e.g., catalogue with all actors involved in a certain LT area, activity with most tools/services or companies, emerging LTs with new resources, etc.)
- training LT-less aware citizens and experts from other communities by including at the LT-centric pages short introductory texts, links to training resources (videos, publications, webinars, etc.), etc.

4 Implementation of the Metadata Model

4.1 Representation principles

The ELG-SHARE Metadata Model is implemented in the form of an XML Schema Definition (XSD) with elements linked to RDF/OWL ontology entities, i.e., each metadata element and value has an identifier which contains the IRI of the corresponding ontology (Figure 5).[40]

```
<xs:element name="LexicalConceptualResource">
    <xs:annotation>
        <xs:documentation xml:lang="en">A resource organised on the basis of lexical or conceptual entries (lexical items,
        terms, concepts etc.) with their supplementary information (e.g., grammatical, semantic, statistical information, etc.)
        </xs:documentation>
        <xs:appinfo>
            <identifier>http://w3id.org/meta-share/meta-share/LexicalConceptualResource</identifier>
            <label xml:lang="en">Lexical/Conceptual resource</label>
            <subclassOf>http://w3id.org/meta-share/meta-share/DataLanguageResource</subclassOf>
        </xs:appinfo>
    </xs:annotation>
```

Figure 5: Example of (part of) an element from the ELG-SHARE Model

This approach presents the following advantages:

- in compliance with the FAIR principles (in particular, Principles I1 – *(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation* and *I2: (Meta)data use vocabularies that follow the FAIR principles*) and the Linked Data paradigm, we use for the ELG-SHARE Model the RDF/OWL language, by including for the metadata elements and values an IRI pointing to the corresponding ontology entity, and support import/export of metadata records in the JSON-LD serialization format

[40] The identifier could also have been used for the element name, which would make the link easier to use for computational systems, but would make the XML records more difficult to create and read for humans.

- the use of XSD enables us to add cardinality to the ontology properties and to represent the metadata contents in the form of an entity-relationship model, facilitating its documentation and conversion into the ELG catalogue backend relational database (see Section 8 of D2.2).

The ELG-SHARE Model utilizes the updated META-SHARE ontology (version 2), and imports the OMTD-SHARE ontology[41] for specific properties and classes (e.g., the LT taxonomy described in the previous section).

The mapping of the ontology entities into XML elements has been automatically performed using a conversion script, which empowers an easy update of the schema alongside the ontology update. Here we should note that where XML distinguishes between *elements*, *attributes* and *values*, RDF/OWL distinguishes between *classes, properties (object* and *data properties*) and *instances.* For the conversion, we have created a set of specifications for their mapping. Some of the main specifications used for the conversion script include, e.g.,:

- mapping of *data properties* into simple *elements* with the corresponding XML *datatype*
- mapping of *classes* with *instances* or (in specific cases, such as the LT taxonomy) *subclasses* into *elements* with *enumerated values* (controlled vocabularies)
- mapping of *classes* that introduce a classification vocabulary or identification scheme into *attributes*
- when there's a pair of *object property* and the *class* it has in its range, selecting to map only one of them into an *element* on the basis of specific criteria.

The structuring of the elements, i.e., attaching the various elements to other elements and appropriately grouping them together, was prepared manually in order to ensure correct attachment and desired ordering. In the ELG-SHARE Model, we have opted for an approach following the Linked Data paradigm instead of the Component Metadata approach, i.e., we attach all *elements – properties* directly to *elements – classes* and refrain from using the extra layer of *wrapper elements* that correspond to the component level. For instance, the "IdentificationInfo" element which was used in META-SHARE for grouping together identifiers, names, descriptions, etc. (i.e., all the elements that can be used for identifying a resource) is not used in ELG-SHARE; instead, all these elements are directly attached to the resource level. This facilitates the exchange of information with other external resources and makes the structure simpler but loses in terms of organizing the elements. Still, for applications where this is needed, the grouping of the metadata elements (*properties* and *classes*) into components and their addition to the model can easily be done.

In addition, the XSD contains a limited set of elements that are used for glueing together the model entities. The root element in the XSD is the *MetadataRecord* that groups together all elements required for describing a metadata record (Figure 6).

The ELG entities that can be described by a metadata record are grouped under a *choice* element *describedEntity* that is used only in the XSD for separating the entities (Figure 7). In a similar way, an element *LRSubclass* attached to the *LanguageResource* is used to group together the LR resource types (e.g., corpus, tool/service, etc.).

[41] Version 2, in progress, http://w3id.org/meta-share/omtd-share/

Figure 6: ELG-SHARE Metadata record

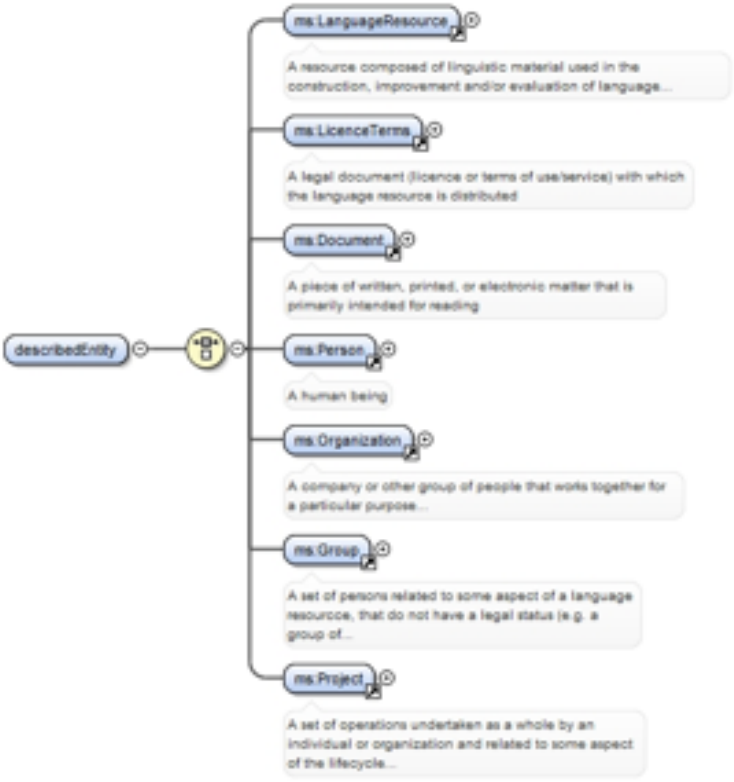

Figure 7: ELG entity described in a metadata record

4.2 Management of metadata records

A core element in the ELG catalogue backend is the REST API for the management of metadata records, i.e., actions for creation, retrieval, update and deletion of metadata records (see Section 8 of D2.2). The metadata management REST API produces and consumes JSON files that contain all the information described in the ELG-SHARE Metadata Model. This REST API will be the entry point between the ELG catalogue backend and the different sources for populating the ELG catalogue, i.e., the ELG metadata editors, the batch ELG compliant metadata import service and the harvesting service on the basis of agreed harvesting protocols. Independent of the source, the provided JSON file must be compliant with the ELG-SHARE model. Validation services will be implemented, and when input JSON files are not compliant, meaningful messages will be provided.

4.3 Interoperability and linking with external resources

One of the objectives of ELG is to bring together resources and information of interest to the LT community; to accomplish this, we intend to deploy Linked Data principles. For this reason, the META-SHARE ontology is in the process of being enriched with semantic relations to other widespread metadata ontologies and vocabularies, including at least the following: Dublin Core[42], DCAT[43], FOAF[44], and Schema.org[45]. These can then be exploited for appropriately exposing the ELG resources to other schemas[46] and catalogues (e.g., Google Dataset search engine[47], LLOD[48], etc.) and in harvesting/converting metadata records from other sources.

5 Future work

Summing up, the remaining work left for the ELG-SHARE Metadata Model includes the following actions:

- improvement of the billing module, following discussions mainly with ELG commercial users,
- getting feedback and incorporating it in the next platform releases,
- finalizing and implementing the relations to other ontologies and vocabularies.

The ELG platform implementation plan includes the relevant actions for creating the support mechanisms for import/export, storage and overall management of the metadata records (e.g., metadata editor, REST API for upload/download, metadata converters, display pages, etc.).

---

[42] https://www.dublincore.org

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

[44] http://xmlns.com/foaf/spec/

[45] https://schema.org

[46] Although it is considered best practice to re-use properties and classes from other well-established ontologies and vocabularies instead of creating new ones, we have decided for this version of the META-SHARE ontology to keep our own concepts and add relations, so that we have better control over them (e.g., adding our own definitions and examples, ensuring that all required information is included and well modelled in the schema) and explore more systematically the most appropriate relations for each of them (i.e., equivalence vs. hierarchical relations).

[47] https://toolbox.google.com/datasetsearch

[48] Linguistic Linked Open Data, http://linguistic-lod.org/llod-cloud

6 References

Broeder, Daan, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, et al. (2008) “Foundation of a Component-Based Flexible Registry for Language Resources and Technology”. In *Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA) <http://www.lrec-conf.org/proceedings/lrec2008/pdf/364_paper.pdf>

Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, et al. (2012) “The META-SHARE Metadata Schema for the Description of Language Resources”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*, Istanbul, Turkey. European Language Resources Association (ELRA) <http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf>

Jörg, Brigitte, Hans Uszkoreit and Alastair Burt (2010) “LT World: Ontology and Reference Information Portal”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010),* Malta. European Language Resources Association (ELRA)

Labropoulou, Penny, Dimitris Galanis, Antonis Lempesis, Mark Greenwood, Petr Knoth, Richard Eckart de Castilho, et al. (2018) “OpenMinTeD: A platform Facilitating Text Mining of Scholarly Content”. In *WOSP 2018 Workshop Proceedings, Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA) <http://lrec-conf.org/workshops/lrec2018/W24/pdf/13_W24.pdf>

McCrae, John, Penny Labropoulou, Jorge Gracia, Marta Villegas, Víctor Rodríguez-Doncel, and Philipp Cimiano (2015) “One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web”. In *The Semantic Web: ESWC 2015 Satellite Events*, ed. by Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, Lecture Notes in Computer Science, pp. 271–82. Springer International Publishing <https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42>

Piperidis, Stelios, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou (2018) “Managing Public Sector Data for Multilingual Applications Development”. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA) <http://www.lrec-conf.org/proceedings/lrec2018/pdf/648.pdf>

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. (2016) “The FAIR Guiding Principles for Scientific Data Management and Stewardship”. *Scientific Data*, 3, 160018 <https://doi.org/10.1038/sdata.2016.18>

A. ELG-SHARE schema documentation on the ELG website

The ELG-SHARE XSD schema is available at:

https://www.european-language-grid.eu/wp-content/uploads/metadata/Schema/ELG-SHARE.xsd

A human-readable documentation in HTML format and links to the sub-schemas listed in the following are available at (see Figure 8):

https://www.european-language-grid.eu/metadata/

- Person
- Organization
- Group
- Licence/Terms
- Project
- Document
- LanguageResource, with common features for all LRs and subschemas for
    - Tool/Service
    - Corpus
    - LanguageDescription
    - Lexical/Conceptual resource.

B. Examples of metadata records

The following example metadata records are provided at https://www.european-language-grid.eu/metadata/.

- Corpus
    - Metadata record for a monolingual corpus (in XML and JSON)
    - Metadata record for a parallel corpus (in XML and JSON)
- Tool/Service
    - Metadata record for Annie NE (in XML and JSON)
    - Metadata record for Cogito Discover Semantic annotator (in XML and JSON)
    - Metadata record for CymrIE (in XML and JSON)
    - Metadata record for Tilde TTS (in XML and JSON)
- Organization
    - Academic institution (in XML and JSON)
    - SME (in XML and JSON)
- Project (in XML and JSON)

Figure 8: The Metadata page on the ELG website at http://www.european-language-grid.eu/metadata/

c. Acknowledgements

The ELG-SHARE Metadata Schema is the outcome of a collaborative effort including contributions from the experts listed below. Given that the schema deploys the updated versions of the META-SHARE and the OMTD-SHARE ontologies, we also list here the main contributors for the original versions.

ELG Metadata Task Force

- Victoria Arranz (ELDA)
- Gerhard Backfried (SAIL LABS)
- Aivars Berzins (TILDE)
- Lucille Blanchard (ELDA)
- Kalina Bontcheva (USFD)
- Khalid Choukri (ELDA)
- Thierry Declerck (DFKI)
- Miltos Deligiannis (ILSP)
- Erinc Dikici (SAIL LABS)
- Ela Elsholz (DFKI)
- Dimitris Galanis (ILSP)
- Andrés Garcia Silva (EXPERT SYSTEM)
- Maria Gavriilidou (ILSP)
- Katerina Gkirtzou (ILSP)
- José Manuel Gómez Pérez (EXPERT SYSTEM)
- Marwa Hadj Salah (ELDA)
- Florian Kintzel (DFKI)
- Penny Labropoulou (ILSP)
- Andis Lagzdins (TILDE)
- Valérie Mapelli (ELDA)
- Katrin Marheinecke (DFKI)
- Juilja Melnika (TILDE)
- Maria Moritz (DFKI)
- Stelios Piperidis (ILSP)
- Herve Pusset (ELDA)
- Georg Rehm (DFKI)
- Ian Roberts (USFD)
- Alexandre Sicard (ELDA)
- Raivis Skadins (TiLDE)
- Severin Stampler (SAIL LABS)
- Andrejs Vasiljevs (TILDE)

External ELG-SHARE collaborators

- Jorge Gracia
- John McCrae
- Víctor Rodríguez Doncel
- Deirdre Lee
- Armando Stellato
- Marta Villegas

META-SHARE ontology contributors

- John P. McCrae
- Jorge Gracia
- Maria Gavriilidou
- Marta Villegas
- Penny Labropoulou
- Philipp Cimiano
- Víctor Rodríguez Doncel

OMTD-SHARE ontology contributors

- Sophie Aubin
- Richard Eckart de Castilho
- Dimitris Galanis
- Katerina Gkirtzou
- Petr Knoth
- Penny Labropoulou
- Claire Nedellec
- Marta Villegas