Chapter 6
Interoperable Metadata Bridges to the wider Language Technology Ecosystem

Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis, Maria Giagkou, Ondřej Košarko, Jan Hajič, and Georg Rehm

Abstract One of the objectives of the European Language Grid is to help overcome the fragmentation of the European Language Technology community by bringing together language resources and technologies, information about them, Language Technology consumers, providers and the wider public. This chapter describes the mechanisms ELG has put in place to build interoperable bridges to related initiatives, infrastructures, platforms and repositories in the wider Language Technology landscape. We focus on the different approaches implemented for the exchange of metadata records about, in a generic sense, resources and exemplify them with the help of four use cases through which the ELG catalogue has been further populated. The chapter presents the protocols used for the population processes as well as the adaptations of the ELG metadata schema and platform policies that proved necessary to be able to ingest these new records. Last, we discuss the challenges emerging in large-scale metadata aggregation processes and propose a number of alternative options to address them.

1 Introduction

One of the objectives of the European Language Grid is to help overcome the fragmentation of the European Language Technology community by bringing together language resources and technologies, information about them, Language Technology consumers, providers and the wider public.

Additionally, ELG is meant to support digital language equality in Europe (STOA 2018; European Parliament 2018), i. e., to create a situation in which *all* European

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

Ondřej Košarko · Jan Hajič
Charles University, Czech Republic, kosarko@ufal.mff.cuni.cz, hajic@ufal.mff.cuni.cz

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

© The Author(s) 2023

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_6

languages are supported through technologies equally well. Technological support for human languages has been characterised by a stark predominance of LTs for English, while almost all other languages are only marginally supported and, thus, in danger of digital extinction (Kornai 2013; Rehm et al. 2014, 2020b; ELRC 2019; Calzolari et al. 2011; Soria et al. 2012). More than ten years after the initial findings (Rehm and Uszkoreit 2012), Europe's languages are still affected by this stark imbalance in 2022, as attested in the most recent series of Language Reports (Giagkou et al. 2022) prepared by the European Language Equality[1] project, which develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030. In collaboration with ELG, one of the first steps towards Digital Language Equality has been the creation of an inventory of language resources and technologies available for Europe's languages and its regular monitoring.

In tandem with its operation as an integrated LT platform, through a battery of selection, conversion and ingestion processes described in this chapter, ELG aims to act as a one-stop shop and single entry point to homogenised descriptions of language resources and technologies. Section 2 positions the ELG approach towards this goal in the broader context of the exchange of metadata between catalogues and repositories. Section 3 presents four use cases through which the ELG catalogue has been populated with metadata records from other sources, highlighting the features that have influenced the different solutions we adopted. Section 4 presents the adaptations made in the ELG metadata schema and platform policies to take into account the outputs of these import procedures. Finally, in Section 5 we discuss, based on the experience gained in this process, the challenges that need to be addressed in the aggregation of metadata from multiple sources in order to share and promote the use and re-use of resources, data and software among community members.

2 Approach

There are a wide range of digital catalogues, repositories and, in general, infrastructures[2] that support the publication and dissemination of digital artefacts and resources, which can be classified along various dimensions. Institutional catalogues hosting all types of resources (publications, datasets, tools, etc.) produced by practitioners affiliated with an institution, catalogues that focus on resources produced by specific communities (e. g., OLAC[3] for resources related to language and linguistics, CLARIN[4] and ELRA[5] for language resources, Europeana[6] for cultural works,

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

[2] For the sake of brevity, we will use the cover term "catalogue" for all institutions of this kind.

[3] http://www.language-archives.org

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

[5] http://elra.info

[6] https://www.europeana.eu

ELIXIR[7] for bioinformatics, LLOD cloud[8] for linguistic linked data, etc.), catalogues that collect specific content types (e. g., Hugging Face[9] for Machine Learning models and datasets, ELRC-SHARE[10] for Machine Translation-related resources or portals for open government data).[11]

At the same time, we witness a strong movement towards the sharing of resources from multiple sources and various disciplines through a common point of access, so that they are easily discoverable, accessible and re-usable by all interested stakeholders, fostering interdisciplinary research and cross-community collaborations as well as Open Science (e. g., European Commission 2022). Google has implemented its Dataset Search[12], a service dedicated to facilitating the discovery of datasets stored across the World Wide Web based on keyword search (Benjelloun et al. 2020). The European Open Science Cloud (EOSC)[13], initiated by the European Commission, is conceived as a federated and open multi-disciplinary environment for hosting and processing research data and all other digital objects produced along the research life cycle, e. g., methods, software and publications (Abramatic et al. 2021). Some European countries have launched corresponding national initiatives, including the National Research Data Infrastructure in Germany (NFDI).[14] Gaia-X[15] seeks to establish a federated ecosystem in which data is made available, collated, shared and processed in trustworthy environments, associated with the concept of data spaces, a type of data relationship between trusted partners, each of whom apply the same high policies, standards and technical components to the description, storage and sharing of their data and other resources.

All these initiatives offer catalogues, or inventories, employing, in many cases, different metadata schemas for the description of resources. The differences between the schemas can be attributed to the varying requirements defined by the relevant object of description (e. g., dataset vs. software or publication or geospatial data), the need to cover a wide range of users (for general catalogues) in contrast to the specialised practices common among scholars of a discipline, as well as to the different purposes that catalogues may serve (e. g., preservation, dissemination, or processing). Sharing metadata across catalogues presupposes interoperability, in particular, *semantic* interoperability. Initiatives for the adoption of common standards in metadata vocabularies, documentation of the vocabularies themselves, and the creation and publication of mappers between them are among the primary instruments to achieve such interoperability (Chan and Zeng 2006; Zeng and Chan 2006; Haslhofer and Klas 2010; Alemu et al. 2012; Broeder et al. 2019).

---

[7] https://elixir-europe.org

[8] https://linguistic-lod.org/llod-cloud

[9] https://huggingface.co

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

[11] https://www.re3data.org/browse/ provides a registry of research data repositories.

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

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

[14] https://www.nfdi.de

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

Equally important is the establishment of protocols and mechanisms for the sharing of metadata, and subsequently of the resources themselves. The OAI-PMH protocol[16] is one of the most popular mechanisms used for repository interoperability at the metadata level. The ResourceSync[17] specification is a framework for the synchronisation of both metadata and resources. Finally, APIs are frequently offered nowadays as a solution for downloading dumps of metadata records.

ELG has established technical bridges with other infrastructures and initiatives in order to enrich its catalogue with information about data resources and tools from other catalogues and repositories. The catalogues of interest to ELG are usually discipline-specific, targeting the LT/NLP and neighbouring areas, such as Machine Learning, Artificial Intelligence as well as social sciences and humanities. Potentially interesting resources for LT development purposes are also hosted in general repositories and catalogues, the identification and filtering of which poses challenges which are briefly discussed in Section 3.

3 Establishing Interoperable Connections: Four Use Cases

Depending on the source repositories' respective contents, metadata schemas and vocabularies, and the available export functionalities of their catalogues, we have adopted different approaches towards establishing interoperable connections, a selection of which is presented in the following use cases. For each use case, we describe the source repository's technical and metadata features, explain how these impact the import of metadata records into ELG and present the methodology and tools used in the integration process.

3.1 Use Case 1: OAI-PMH (CLARIN Nodes and ELRC-SHARE)

The CLARIN (Common Language Resources and Technology Infrastructure) Research Infrastructure (Hinrichs and Krauwer 2014; Eskevich et al. 2020) supports the sharing, use and sustainability of digital language resources and tools for research in the social sciences and humanities. It is established in the form of a networked federation of centres (Wittenburg et al. 2010), consisting of language data repositories, service centres and knowledge centres, with single sign-on access for all members of the academic community in all participating countries.

As part of the technical interoperability specifications, CLARIN data repositories are required to expose their metadata records to the Virtual Language Observatory[18] using OAI-PMH. With regard to metadata interoperability, CLARIN has designed

[16] https://www.openarchives.org/pmh/

[17] http://www.openarchives.org/rs/1.1/resourcesync

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

and implemented the Component MetaData Infrastructure (CMDI)[19], a framework for the description and reuse of metadata "components" (semantic groups of elements) which can be combined to build "profiles", i. e., metadata templates for specific resource types by specific communities or groups (Broeder et al. 2008, 2012). Both are stored and shared through a dedicated registry, with metadata records being shared in the form of XML files compatible with one of these profiles.

The ELG platform implements an OAI-PMH client for harvesting metadata from external repositories which expose their metadata via OAI-PMH. The process of harvesting requires the registration of a third-party provider as an "OAI-PMH Provider" in the ELG catalogue. As soon as communication is established, the third-party provider shares their OAI-PMH endpoint, which ELG will call at regular intervals (currently once a week) in order to harvest the metadata the external repository exposes. Thus, for linking with the CLARIN infrastructure, the OAI-PMH harvesting protocol is the ideal candidate.

The metadata schema is a crucial parameter to be taken into account in the harvesting process. The ELG harvester accepts metadata records compliant with the minimal version of the ELG metadata schema (see Section 5 in Chapter 2). LINDAT/CLARIAH-CZ[20], the Czech CLARIN national node, does indeed expose its metadata records described using the META-SHARE minimal schema through its OAI-PMH endpoint (Gavrilidou et al. 2012). The fact that the ELG schema (Labropoulou et al. 2020) builds upon META-SHARE proved valuable in the conversion process of the original LINDAT/CLARIAH-CZ metadata into the ELG schema (see Chapter 8, Section 4, p. 157 ff., for more technical details).

CLARIN-DSpace, the repository software[21] (forked from DSpace[22]) developed mainly by the LINDAT/CLARIAH-CZ team, is used by several CLARIN centres for their repositories (Straňák et al. 2019). After pulling the latest changes, these repositories are ready-to-import into ELG using the same harvesting mechanism and procedure. At the time of writing, the mechanism described above is also used for harvesting CLARIN-PL[23] and CLARIN-SI[24].

The same harvesting approach was followed for the harvesting of metadata records from the ELRC-SHARE repository, which is used for the storage of and access to language resources collected through the European Language Resource Coordination[25] initiative (Lösch et al. 2018) and for feeding the CEF Automated Translation (CEF.AT) platform.[26] ELRC-SHARE (Piperidis et al. 2018) uses a metadata schema based on the META-SHARE schema tuned to text resources for Machine

[19] https://www.clarin.eu/content/component-metadata

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

[21] https://github.com/ufal/clarin-dspace

[22] https://duraspace.org/dspace/

[23] https://clarin-pl.eu/dspace/

[24] https://www.clarin.si/repository/xmlui/?locale-attribute=en

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

[26] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

Translation purposes. Again, the mapping of the metadata records from the original schema to ELG was undertaken by the two teams.

3.2 Use Case 2: Custom API and Proprietary Schema (Hugging Face)

A different procedure is used for catalogues that expose metadata records through custom APIs and proprietary metadata schemas. This procedure is used only for catalogues that are of high interest to the ELG objectives. The Hugging Face catalogue (Wolf et al. 2020) is such a case. It is a large collection of machine learning models and datasets that can be used for training models, with a focus on the Transformer architecture. Since 2021 ELG and Hugging Face have been collaborating with the goal of importing metadata records from the Hugging Face catalogue into ELG.

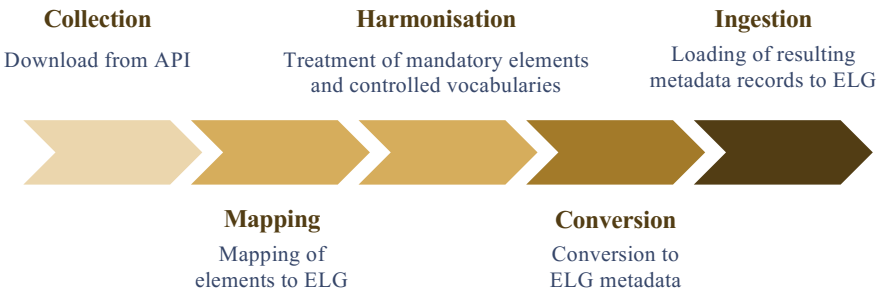

Fig. 1 Workflow for the import of metadata records from Hugging Face to ELG

One of the goals of Hugging Face is to enable its users to upload datasets and models following a set of specifications so that they can be deployed for testing and building other models or integrating models in their applications. Although they encourage users to add descriptions for the resources, this is not enforced. Furthermore, the suggested metadata elements do not follow a standard schema. Users are asked to upload a "card" for datasets[27] or models[28], with a combination of free text fields and a set of tags (e. g., language, licence) with values from recommended controlled vocabularies, which are, however, not strictly validated.

Hugging Face exposes two APIs with JSON files for datasets and models respectively. These JSON files include a subset of the metadata elements displayed in their catalogue, however, not all records have values for all of the elements. Since importing into ELG presupposes that the metadata records comply with the ELG metadata schema, which means that at least the mandatory elements of the minimal version (see Section 5 in Chapter 2) are filled in, the conversion and import of records from Hugging Face into ELG has so far been limited to datasets with at least the de-

[27] https://huggingface.co/docs/datasets/dataset_card.html

[28] https://huggingface.co/docs/hub/model-repos

scription, language and licence elements filled in as these are deemed the minimum threshold for findability and usability purposes in the context of ELG.

A conversion process has been set up based on the mapping of the elements and, in the case of controlled vocabularies, their values. Further enrichment of the resulting records has been performed for specific elements. The most prominent case was that of the licencing information, since ELG requires, besides its name, a URL with the text of the licence. Hugging Face includes a list of licence identifiers taken from the SPDX list[29] (which are also used in ELG), but it allows users as well to add a licence name without further information. Thus, in addition to the mappings of the licence identifiers from Hugging Face into the ones used in ELG, we looked for the licence URL of unmapped values; if no URL was found, the resource was not imported into ELG. Finally, where required, default values have been used for mandatory elements whose values could not be inferred from the original metadata records (e. g., all datasets have been assigned the text value for media type). Figure 1 shows the workflow that was followed in this process.

3.3 Use Case 3: General Catalogues and Standard Schemas (Zenodo)

Catalogues with heterogeneous resources from multiple sources and disciplines present various challenges. We use Zenodo[30] to discuss these challenges.

Zenodo[31] is a repository for storing and sharing EC-funded research results to support Open Science established and run by CERN, which was created in response to the European Commission's (EC) assignment to the OpenAIRE project.[32]. Since its launch, Zenodo has grown steadily and is currently used for the publication of all types of resources beyond EC-funded ones by research communities and individuals. The constant update of the Zenodo catalogue and its uptake by researchers for the upload of datasets, and, more recently, software, makes it particularly interesting for ELG purposes. The size and increasing number, however, of catalogue contents makes the selection of resources very challenging. During the first phase of the ELG project, we used a manual process for the identification of resources, which is described in Chapter 8. This process, though, does not allow for regular updates and has been abandoned in favour of an automatic process.

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

[30] https://zenodo.org

[31] https://about.zenodo.org

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

Zenodo exposes its metadata records through two channels: a REST API[33], which outputs records as JSON files, and an OAI-PMH API[34] in a set of standard metadata formats, i. e., DC[35], DataCite[36], MARC21[37] and DCAT[38].

With regard to the ELG import mechanism, our preferred solution is OAI-PMH, a standard protocol for interoperability and exchange of metadata records, which includes a mechanism for regular harvesting. However, the Zenodo OAI-PMH endpoint does not allow the selection based on resource types, which would allow us to focus on "datasets" and "software". The only option is to download the whole set of metadata records in order to subsequently filter them. Furthermore, harvesting from the OAI-PMH endpoint is rate limited, hence not appropriate for large numbers of metadata records. We have, therefore, resorted to a combined solution:

- We downloaded a full dump of 2,060,674 metadata records included in Zenodo up until 31 August 2021. This dump, which is available from Zenodo, contains all records in JSON format, was filtered according to resource-type.
- For records added to Zenodo after this date, we are incrementally harvesting from the OAI-PMH endpoint. Through this channel, a set of additional 147,621 records has been harvested in a three-month period.

The next step is that of identifying the candidate resources for ELG. From the 2,208,295 metadata records available up until 31 December 2021, those of resource type "dataset" and "software" amount to 592,509 entries. This number is rather high, and since the majority of these records are of little or no interest to ELG users[39], we are experimenting with automated filtering methods to identify the records of interest.

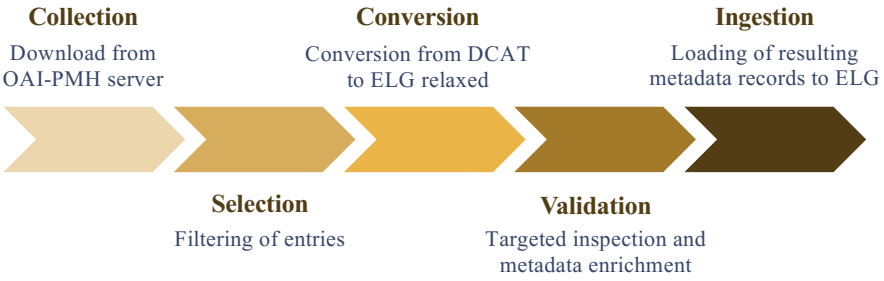

Fig. 2 Workflow for the import of metadata records from Zenodo to ELG

[33] https://developers.zenodo.org/#rest-api
[34] https://developers.zenodo.org/#oai-pmh
[35] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[36] https://schema.datacite.org/meta/kernel-4.4/
[37] https://www.loc.gov/marc/bibliographic/
[38] https://www.w3.org/TR/vocab-dcat-3/
[39] As a comparison, the ELG catalogue has approx. 13,000 metadata records at the time of writing.

The conversion of the metadata records is based on the DCAT metadata schema (Albertoni et al. 2022), which is in widespread use. We expect that mapping DCAT to ELG will enable the re-use of these converters as a base for import from other repositories. Moreover, DCAT is the schema with the richest information among the ones exposed from Zenodo, and the only one that includes a direct link to the downloadable files ("downloadURL" element), an important feature for ELG consumers.

Mapping from DCAT is, however, not straightforward. DCAT is an RDF vocabulary, and restrictions and extensions are implemented in the form of profiles and applications. The OAI-PMH endpoint makes the metadata records available in XML format; the XSD schema used by Zenodo is not publicly available[40]. A closer inspection of the XML files has revealed discrepancies in the representation of some elements. For instance, "subject" (defined in DCAT as a SKOS[41] Concept) appears in Zenodo XML files either as a SKOS Concept or as an element with the IRI of the subject value in the form of an attribute. We have analysed the Zenodo XML files, to the extent possible, and based our mapping on this analysis. We also had to apply some modifications in the ELG schema so that we could take into account the DCAT features (Section 4.1). Finally, a converter for the elements in the JSON files offered through the REST API for the first batch of files has also been implemented.

As a result of this endeavour, the procedure for regular updates from Zenodo is foreseen as a workflow integrating the following steps: harvesting from the Zenodo endpoint, offline filtering and conversion of the metadata records, possibly with some manual targeted inspection, and import into ELG (Figure 2).

3.4 Use Case 4: Collaborative Community Initiatives (ELE, ELG)

We also populated the ELG catalogue using bulk lists of metadata records, potentially containing limited information, that serve as seeds for further enrichment. We present here two such cases, one set of resources collected collaboratively in ELE and a second set collected by the ELG consortium.

The European Language Equality (ELE) project (Rehm and Way 2023)[42], which collaborates with ELG to promote digital language equality in Europe, launched a project-internal initiative in 2021 to collect as many LRTs as possible available for the languages under investigation by the project.[43] Operationally, a web form was set up, which included a subset of the mandatory metadata elements of the ELG schema. Given the size and breadth of this activity (dozens of respondents throughout Europe for approx. 80 official, regional, minority languages), we considered requiring every informant to fill in even the minimal version of the metadata schema for every single resource identified too demanding and not paricularly realistic, perhaps

[40] The XSD schema included in the OAI-PMH API for DCAT is in fact that of DataCite v4.1.

[41] https://www.w3.org/2004/02/skos/

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

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

even negatively impacting the collection process itself, potentially resulting in fewer resources being reported by the informants if the process of registering a resource took too much time. The modifications required to accommodate this collaborative scenario resulted in a "relaxed" version of the schema (see Section 4.1).

The results of this collection process were exported in a tabular format. Before the conversion and final import of the approx. 6,500 records into ELG, a long and demanding process of curation was undertaken using semi-automatic methods. The final output was imported into ELG through various scripts (Figure 3).

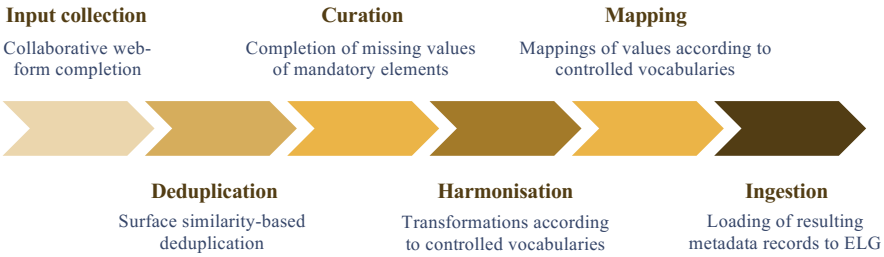

Fig. 3 Workflow for the import of ELE results to ELG

The curation process included normalising, correcting, and enriching values of elements that were absent or not used consistently. Despite the effort to control the input through prompting for the selection of values from recommended vocabularies and filling in mandatory values, web forms do not allow strict enforcement strategies, especially for cases of long lists of values or multiple values. For example, although a set of "language" values was offered for selection in the form, the informants could also add other values, which resulted in values with alternative, unofficial or simply unusual names. Therefore, language information had to be normalised and mapped to the ISO 639 language codes, as required by ELG. Although the tabular format presents some advantages, given its simplicity and users' familiarity, it still poses a number of challenges for validation purposes, especially for elements with patterns, or with multiple values. For instance, the "email" element was filled in with free text values, URL links, etc., since no validation pattern was used for the element. For elements with multiple values, such as languages, functions, etc., different delimiters were used in between values and had to be normalised. Moreover, nested information cannot be represented in a flat form; for example, the values of language and region (where the language is spoken) were split in two complementary columns so that controlled vocabularies could be used, but there can be no guarantee that both columns are consistently filled in. For these cases, we had to check and ensure that the same number of values was consistently used across the two complementary columns and, moreover, that the values were matched correctly.

In a similar collaborative population setting, the catalogue was populated with European organisations that develop or use LTs or LRs, which were collected by the ELG team and the National Competence Centres (NCCs; see Chapter 11 for more details), thus enabling ELG to quickly become the "yellow pages" of organisations

active in the broader LT community. As described in more detail in Chapter 9, lists of organisations from various sources have been merged, together with information on list items – mainly contact data and key terms describing their LT-related activities. The resulting enriched list, divided into sub-lists by country, was checked again by the respective NCCs, and, after checking the consistency, more than 1,700 records were converted into the ELG-compatible XML format and imported into ELG. At the time of writing, a similar procedure is being followed for LT-related R&D projects and their funding agencies.

3.5 Summary of Use Cases

Table 1 summarises the technical and the metadata conditions in each of the use cases presented in this section and the ways these are catered for in ELG. Depending on the export functionalities offered by the source, the ELG platform can establish a connection at regular intervals and benefit from continuous updates. Table 1 also shows the ELG metadata schema version that can be used, depending on the source metadata schema, as well as the quantity and information richness of metadata records.

Repository	Export Functionality	Metadata Schema	ELG Schema Version	Update Frequency
CLARIN nodes	OAI-PMH	META-SHARE	minimal	regular
ELRC-SHARE	OAI-PMH	ELRC-SHARE	minimal	regular
Hugging Face	REST API	Proprietary (JSON)	relaxed	one-off
Zenodo	REST API	Proprietary (JSON)	relaxed	one-off
Zenodo	OAI-PMH	DCAT (XML)	relaxed	regular
ELE survey	–	Subset of ELG schema	relaxed	one-off
ELG collection	–	Subset of ELG schema	relaxed	one-off

Table 1 Overview of use cases

4 Implementing Metadata Interoperability

Primarily motivated by our various interoperability use cases, some of which are described in Section 3, we modified the ELG platform import procedures and policies, especially with regard to the metadata schema and the publication life cycle (described in Chapter 2), so that they are able to handle the different interoperability scenarios. These adaptations are not restricted to the requirements of the use cases but lay the foundational principles for accommodating a broader range of metadata import scenarios.

4.1 ELG Metadata Schema – Relaxed Version

The "relaxed" version of the ELG metadata schema aims to accommodate mismatches between the ELG schema and schemas used for metadata records that are automatically imported into the ELG catalogue, especially those from catalogues with limited information or catalogues populated with metadata records of interest to a broader range of communities (e. g., Zenodo, EOSC, etc.) and, thus, using more general schemas, e. g., DCAT (Albertoni et al. 2022) or DataCite[44] (DataCite Metadata Working Group 2021). This version of the schema features additional alternative elements for mandatory metadata elements that may be missing from the source records or that have different data types.

The first case refers to two elements that are deemed important for ELG purposes: "media type" and "licence".

- The element "media type part" is crucial for ELG, as it is used for attaching important metadata properties, such as language, format, size, etc. Even in cases where these are included in source records, they may come with *different* classification vocabularies and semantics and, therefore, cannot be imported into ELG. For these cases, the additional alternative value "unspecified media part" can be used.
- The element "licence" is crucial for re-usability purposes; for a licence, both a name and a URL hyperlink to the respective legal document are required. However, in many cases, such as legacy resources, or records in catalogues allowing free text as the value of "licence", the name and URL cannot be determined automatically. This is why we introduced the "access rights" element that takes a free text value as an alternative to "licence", specifying the rights of access and use at a higher level of abstraction.

The second case groups together elements which take a value from controlled vocabularies in ELG, while in other schemas they have a free text value (e. g., "service function", "size unit", etc.) and combined elements that cannot be distinguished from the source metadata record (e. g., when size is encoded as free text combining amount and size unit together). To address the first case, we modified the data type of the element so that it takes a value from a recommended vocabulary or free text entered by the user; to address the second case, we introduced a new element that takes free text as a value (e. g., "sizeText" can be used as an alternative to the combination of "amount" and "size unit").

4.2 Publication Policies for Imported Metadata Records

ELG rates the quality of the metadata records highly. High quality metadata contributes to the discovery and usage of the resources themselves. A standardised pub-

[44] https://schema.datacite.org

lication life cycle has been established in ELG for metadata records (see Chapter 2, Section 6, 24 ff.). However, the same level of quality cannot be enforced across all metadata records. This is also taken into account in the publication policies. Thus, while metadata records registered by individuals go through a validation process, for records automatically imported from other catalogues the same manual validation processes cannot be set up in a feasible way, i. e., the quality and extent, in terms of information, of external metadata records remains under the responsibility of the respective source catalogue. Depending on the harvesting process and source catalogue, a three-level classification of metadata records is used:

- *Metadata records harvested automatically from collaborating catalogues (CLARIN nodes, ELRC-SHARE)*, which have similar metadata requirements as ELG. These records are added by individuals, the resource is stored in the repository. This is why these metadata records are considered trustworthy, and the records are published in the ELG catalogue as is, i. e., without any human validation.
- *Metadata records automatically imported from catalogues with "lighter" metadata requirements (Hugging Face, Zenodo)* have originally been added to the source catalogue by individuals together with the physical resource. The metadata record and resource is considered trustworthy but it may lack information which is important for ELG purposes, and thus marked as "for information" to indicate to ELG users that important information may be missing.
- *Metadata records that resulted from bulk collection initiatives (ELE collection, ELG collection)* are often incomplete, i. e., only a subset of the required information was collected and converted to the ELG schema. These records adhere to the relaxed ELG schema, the physical resource may be stored anywhere online. These records do not undergo the validation process, they are marked and can be claimed for further enrichment by their rightful owners (see Chapter 9, Section 3.3, p. 179). When a user claims a metadata record, the technical ELG team is notified and can approve or reject the claim, taking into account the professional email account of the user; if the claim is approved, the metadata record is unpublished and assigned to the user for further editing. Once the user finishes the editing, the record is submitted for publication and goes through the normal publication procedure. Users are notified about the claim procedure of these metadata records via e-mail.

5 Interoperability across Repositories

The interoperability across multiple repositories and platforms is of utmost importance in a broader, federated environment of data and services, as envisaged in initiatives like EOSC (European Open Science Cloud, see, e. g., Corcho et al. 2021), NFDI, Gaia-X or the European Commission's Data Spaces and in accordance with the FAIR principles (Wilkinson et al. 2016), see Section 2. In the following, we discuss some of the open issues that need to be addressed in order to achieve this based on the endeavours presented in this chapter.

5.1 Technical Interoperability across Repositories

The first prerequisite for the sharing of metadata records and the construction of a common master inventory based on the contents of all participating repositories is that of exchange services. The OAI-PMH protocol, despite its limitation to the exchange of metadata, constitutes the most widespread and hence usually preferred option. REST services are becoming more popular, but they are not yet standardised and thus require customised solutions. Rehm et al. (2020a) explore technical and semantic interoperability in more detail.

5.2 Semantic Interoperability across Repositories

The use of shared vocabularies for the documentation of resources is the next necessary step towards interoperability. The standardisation and documentation of metadata schemas is a requirement that many initiatives have articulated (Hugo et al. 2020; Behnke et al. 2021). While certain metadata vocabularies, such as DC[45], DCAT, schema.org[46] and DataCite, have become de facto standards, these are general schemas that can be used to express core metadata elements required for the description of any type of digital resource. This, however, competes with the much more fine-grained documentation needs of specific communities and more detailed requirements set to achieve machine actionability. For example, "resource type" is an element that poses problems for all catalogues: in contrast to the general vocabularies (e. g., COAR resource type vocabulary[47], a limited set of values from DC[48], Zenodo[49]), communities prefer finer distinctions (cf. the values of "resource type" in the CLARIN VLO[50]). This creates a burden when moving from general to specialised catalogues (e. g., from Zenodo to ELG).

Bridges and mappers between vocabularies are developed, especially between the popular schemas.[51] Yet this is not a scalable approach, as for each new vocabulary a new mapper has to be built. Instead, a "shared semantic space" is needed as a joint, ontologically grounded and machine-readable vocabulary, into which all concepts and terminologies can be mapped (Rehm et al. 2020a). This space can be envisaged as a reference model able to represent all crucial information typically contained in the respective metadata schema. However, a single RDF/OWL ontology covering general and domain or community-specific semantic categories is an almost impossible task to achieve (Labropoulou et al. 2018). An alternative could be a Linked

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

[46] https://schema.org

[47] https://vocabularies.coar-repositories.org/resource_types/

[48] https://www.dublincore.org/specifications/dublin-core/resource-typelist/

[49] https://developers.zenodo.org/#representation

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

[51] For the mapping of metadata schemas in the wider LT ecosystem, see McCrae et al. (2015b,a).

Data approach[52], in which different communities maintain their independent formal models and vocabularies and subsequently refer to reference vocabularies or concepts developed in a distributed fashion by the broader community. As an example of such an approach, a collaboration was initiated between ELG and the AI4EU project on the mapping of the ontologies used in the two platforms. This work is continued under the umbrella of the AI Ontology Working Group which includes members from the European AI on Demand Platform and collaborating projects.[53]

Even in this scenario, though, an important issue to be addressed is that of the appropriate semantic relations. Equivalence relations are not always one-to-one and also need to take into account the type of elements. Additionally, there are an abundance of similar vocabularies recommended by different communities or serving different documentation needs. For example, in terms of "language", a value taken from ISO 639[54] may suffice for general catalogues. But for the metadata of resources in language-related catalogues, such as ELG, a more detailed value space is required, that takes into account dialects and other varieties, and these are not included in ISO 639 (Gillis-Webber and Tittel 2019). In ELG we use the BCP 47 recommendation (Phillips and Davis 2009) alongside values taken from the Glottolog[55] vocabulary (Hammarström et al. 2021) so that we can exploit the finer distinctions made in it for language varieties. The fact that Glottolog includes a mapping to ISO 639-3 values, when these exist, facilitates this endeavour and the exchange of metadata records with catalogues that prefer using ISO 639.

5.3 Minimal Metadata Requirements

The different purposes served by the catalogues have an impact on the exchange of metadata records, too. For example, Zenodo is used for the publication of research outcomes by many different organisations and individuals. The fact that there is a very small set of mandatory elements as well as the fact that providers do not have a strong incentive to make their resources findable lowers the quality of the metadata descriptions. In a similar way, individuals that add their resources to the Hugging Face catalogue are mostly interested in testing their dataset and do not pay attention to its description. Many metadata elements that are important for ELG purposes, such as "language", are simply not included in the formal descriptions of these records. Often, even free text descriptions are of very low quality and cannot be used for discovery purposes. There is, therefore, a strong need for training resource owners on the importance of metadata together with the continuous curation by experts (Gordon and Habermann 2019). The "claim" procedure adopted in ELG is a step along these lines. Semi-automatic methods for enriching metadata records by extracting

[52] https://www.w3.org/DesignIssues/LinkedData.html

[53] https://www.ai4europe.eu/ai-community/working-groups-d/ontology

[54] https://www.iso.org/iso-639-language-codes.html

[55] https://glottolog.org

information from the datasets themselves, as well as other sources, will also play an important role in ensuring that minimal documentation requirements are met.

5.4 Duplicate Resources

Looking at the resources themselves, the exchange of metadata records across catalogues comes with the risk of creating duplicates and near-duplicates. The same resource may appear with slightly different names in catalogues and similar descriptions, while the same name is often used for subsets of the resource. The use of persistent identifiers (PIDs) has been proposed to address this, but it cannot be guaranteed that persistent identifiers are indeed unique. Explicit relations between similar resources (subsets, raw or annotated versions, versions and updates, etc.) must be formally recorded in the metadata so that they can be used for deduplication purposes. Establishing relations between the metadata records of the same resource in different catalogues should also be recorded.

6 Conclusions

In this chapter we have focused on the sharing of metadata between catalogues. This is only the basis for what is going to be the next level of sharing data and software which is the ultimate goal. This involves not only a shared semantic space to anchor and cross-link metadata vocabularies but also technical compatibility and cooperation. ELG has closely collaborated with other platforms to explore platform interoperability at various levels (Rehm et al. 2020a). Experiments were conducted with AI4EU[56], SPEAKER[57] and QURATOR[58] for the creation of cross-platform workflows, where data and services were accessed from one platform and either transferred to another platform or used for building a pipeline or workflow of different processing services in another platform. Our initial experiments, explored further by Moreno-Schneider et al. (2022), demonstrate that interoperability can be partially achieved, with a certain degree of manual and automatic interventions.

Finally, we should also mention an alternative that can be used for sharing resources and their documentations across platforms and communities. This consists of supporting cross-platform search through making search and discovery APIs used by a platform available to third parties so that they can integrate them in their own search space (Rehm et al. 2020a). This way, a single query would return matches from multiple platforms whose publicly available search APIs are integrated in the platform queried by the user. In this case, search results would show only a minimal

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

[57] https://www.speaker.fraunhofer.de

[58] https://qurator.ai

set of metadata redirecting the user to the platform that offers the respective resource. Again, a shared common space is required but only for a limited set of metadata – a similar situation to the general catalogues presented above, but only for a small subset. However, this option presents a scalability problem as soon as the number of collaborating platforms and respective search APIs grows.

Decentralised infrastructures such as Gaia-X, in which individual trusted platforms follow a common standard (i. e., the Gaia-X federation services) and become a networked system freely sharing and exchanging data and services across multiple actors, offer a viable solution addressing this challenge. OpenGPT-X[59] is a German national project in which large language models are currently being developed, especially for German but also for English and other European languages. In this project, which has started in January 2022, we will have the chance to implement the emerging Gaia-X specifications in the ELG platform so that it joins this emerging ecosystem.

References

Abramatic, Jean-François, Jan Hrušák, and Sarah Jones, eds. (2021). *European Open Science Cloud (EOSC) Executive Board: Final Progress Report*. Publications Office. DOI: 10.2777/46019.

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez-Beltran, Andrea Perego, and Peter Winstanley, eds. (2022). *Data Catalog Vocabulary (DCAT) – Version 3*. W3C Working Draft. URL: https://www.w3.org/TR/vocab-dcat-3/.

Alemu, Getaneh, Brett Stevens, and Penny Ross (2012). "Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: a social constructivist approach". In: *New Library World* 113.1/2, p. 15.

Behnke, Claudia, Kees Burger, Yann le Franc, Wim Hugo, Pekka Järveläinen, Jessica Parland-von Essen, and Gerard Coen (2021). "D2.6 First reference implementation of the data repositories features". In: DOI: 10.5281/zenodo.5362027. URL: https://zenodo.org/record/5362027/export/hx.

Benjelloun, Omar, Shiyu Chen, and Natasha Noy (2020). "Google Dataset Search by the Numbers". In: *The Semantic Web (ISWC 2020) – 19th International Semantic Web Conference*. Ed. by Jeff Z. Pan, Valentina A. M. Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal. Vol. 12507. Lecture Notes in Computer Science. Athens, Greece: Springer, pp. 667–682. DOI: 10.1007/978-3-030-62466-8_41. URL: https://doi.org/10.1007/978-3-030-62466-8_41.

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

Broeder, Daan, Thorsten Trippel, Emiliano Degl'Innocenti, Roberta Giacomi, Maurizio Sanesi, Mari Kleemola, Katja Moilanen, Henri Ala-Lahti, Caspar Jordan, Iris Alfredsson, Hervé L'Hours, and Matej Ďurčo (2019). "SSHOC D3.1 Report on SSHOC (meta)data interoperability problems". In: DOI: 10.5281/ZENODO.3569868. URL: https://zenodo.org/record/3569868.

Broeder, Daan, Dieter van Uytvanck, Maria Gavrilidou, Thorsten Trippel, and Menzo Windhouwer (2012). "Standardizing a Component Metadata Infrastructure". In: *Proceedings of the Eighth In-*

[59] https://opengpt-x.de

ternational Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turkey: ELRA, pp. 1387–1390. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/581_Paper.pdf.

Calzolari, Nicoletta, Valeria Quochi, and Claudia Soria, eds. (2011). *The Strategic Language Resource Agenda*. URL: https://www.academia.edu/1651334/The_Strategic_Language_Resource_Agenda.

Chan, Lois Mai and Marcia Lei Zeng (2006). "Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level". In: *D-Lib Magazine* 12.6. DOI: 10.1045/june2006-chan. URL: http://www.dlib.org/dlib/june06/chan/06chan.html.

Corcho, Oscar, Magnus Eriksson, Krzysztof Kurowski, Milan Ojsteršek, Christine Choirat, Mark van de Sanden, Frederik Coppens, and European Commission, Directorate-General for Research and Innovation (2021). *EOSC Interoperability Framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Publications Office. DOI: 10.2777/620649. URL: https://data.europa.eu/doi/10.2777/620649.

DataCite Metadata Working Group (2021). "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". In: DOI: 10.14454/3W3Z-SA82. URL: https://schema.datacite.org/meta/kernel-4.4/.

ELRC (2019). *ELRC White Paper: Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe*. Second online edition. URL: https://lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf.

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

European Commission (2022). *European Research Area policy agenda: overview of actions for the period 2022–2024*. Publications Office. DOI: 10.2777/52110. URL: https://data.europa.eu/doi/10.2777/52110.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. URL: http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

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

Giagkou, Maria, Stelios Piperidis, Georg Rehm, and Jane Dunne, eds. (2022). *Language Technology Support of Europe's Languages in 2020/2021*. Various project deliverables (language reports); EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE. European Language Equality Project. URL: https://european-language-equality.eu/deliverables/.

Gillis-Webber, Frances and Sabine Tittel (2019). "The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages". In: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Ed. by Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski. Vol. 70. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 4:1–4:15. DOI: 10.4230/OASIcs.LDK.2019.4. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10368.

Gordon, Sean and Ted Habermann (2019). *Visualizing The Evolution of Metadata*. Version Number: v0.0.1. DOI: 10.5281/zenodo.2538983. URL: https://doi.org/10.5281/zenodo.2538983.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank (2021). *Glottolog database 4.5*. Version Number: v4.5, Type: dataset. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/ZENODO.5772642. URL: https://zenodo.org/record/5772642.

Haslhofer, Bernhard and Wolfgang Klas (2010). "A survey of techniques for achieving metadata interoperability". In: *ACM Computing Surveys* 42.2, pp. 1–37. DOI: 10.1145/1667062.1667064. URL: https://dl.acm.org/doi/10.1145/1667062.1667064.

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

Hugo, Wim, Yann Le Franc, Gerard Coen, Jessica Parland-von Essen, and Luiz Bonino (2020). "D2.5 FAIR Semantics Recommendations Second Iteration". In: DOI: 10.5281/zenodo.5362010. URL: https://zenodo.org/record/5362010.

Kornai, Andras (2013). "Digital Language Death". In: *PLoS ONE* 8.10. DOI: 10.1371/journal.pone.0077056. URL: https://doi.org/10.1371/journal.pone.0077056.

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

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

Lösch, Andrea, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith (2018). "European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management". In: *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA. URL: https://aclanthology.org/L18-1213.

McCrae, John Philip, Philipp Cimiano, Victor Rodriguez-Doncel, Daniel Vila Suero, Jorge Gracia, Luca Matteis, Roberto Navigli, Andrejs Abele, Gabriela Vulcu, and Paul Buitelaar (2015a). "Reconciling Heterogeneous Descriptions of Language Resources". In: *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*. Beijing, China: ACL, pp. 39–48. DOI: 10.18653/v1/W15-4205. URL: http://aclweb.org/anthology/W15-4205.

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

Moreno-Schneider, Julián, Rémi Calizzano, Florian Kintzel, Georg Rehm, Dimitris Galanis, and Ian Roberts (2022). "Towards Practical Semantic Interoperability in NLP Platforms". In: *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA 2022; co-located with LREC 2022)*. Ed. by Harry Bunt. Marseille, France, pp. 118–126. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/ISA-18/pdf/2022.isa18-1.16.pdf.

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

Piperidis, Stelios, 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)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/648.pdf.

Rehm, Georg, Dimitrios Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julián Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John P. McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš (2020a). "Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability". In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France, pp. 96–107. URL: https://www.aclweb.org/anthology/2020.iwltp-1.15.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020b). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Váradi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad (2014). "An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"". In: *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*. Ed. by Laurette Pretorius, Claudia Soria, and Paola Baroni. Reykjavik, Iceland, pp. 30–37. URL: http://georg-re.hm/pdf/CCURL-2014-META-NET.pdf.

Rehm, Georg and Andy Way, eds. (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Forthcoming. Springer.

Soria, Claudia, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Nicoletta Calzolari (2012). "The FLaReNet Strategic Language Resource Agenda". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 1379–1386. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/777_Paper.pdf.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. URL: https://data.europa.eu/doi/10.2861/136527.

Straňák, Pavel, Ondřej Košarko, and Jozef Mišutka (2019). "CLARIN-DSpace repository at LINDAT/CLARIN : LINDAT/CLARIN FAIR repository for language data". In: *The grey Journal – International Journal on Grey Literature* 16, pp. 52–61.

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

Wittenburg, Peter, Nuria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicova, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco van Veenendaal, Tamas Váradi, and Martin Wynne (2010). "Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: ELRA. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/679_Paper.pdf.

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

Zeng, Marcia Lei and Lois Mai Chan (2006). "Metadata Interoperability and Standardization - A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels". In: *D-Lib Magazine* 12.6. DOI: 10.1045/june2006-zeng. URL: http://www.dlib.org/dlib/june06/zeng/06zeng.html.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.