Noname manuscript No.
(will be inserted by the editor)

Discovering language resources using Linghub

John P. McCrae · Markus Ackermann · Philipp Cimiano · Martin Brümmer · Victor Rodríguez Doncel · Daniel Vila · Gabriela Vulcu · Andrejs Abele · Paul Buitelaar · Luca Matteis · Tiziano Flati · Jorge Gracia

Received: date / Accepted: date

Abstract For most applications of natural language processing, the discovery of suitable resources for a given domain and language is vital, yet current catalogues of language resources are incomplete and vary in the level of detail. We approach this problem by means of using linked data principles to harmonize these heterogeneous resources into a single portal called Linghub. We use state-of-the-art entity disambiguation techniques to combine these entities and provide a single interface to the user. We evaluate this approach by using real user queries harvested from the Corpora Mailing List and show that the linked data technology embedded in this portal can nearly double the precision in finding answers from 18.8% to 35.9% while still increasing the overall number of resource found by 25.3%

John P. McCrae, Gabriela Vulcu, Andrejs Abele, Paul Buitelaar
Insight Centre for Data Analytics
National University of Ireland Galway
E-mail: john@mccr.ae, {gabriela.vulcu, andrejs.abele}@insight-center.org

Markus Ackermann
Institute for Applied Informatics
Leipzig University
E-mail: ackermann@informatik.uni-leipzig.de

Víctor Rodríguez Doncel, Jorge Gracia
Ontology Engineering Group
Polytechnic University of Madrid
E-mail: {vrodriguez, jgracia}@fi.upm.es

Luca Matteis, Tiziano Flati
University of Rome "La Sapienza"
E-mail: {matteis, flati}@di.uniroma1.it

John P. McCrae, Philipp Cimiano
Cognitive Interaction Technology, Excellence Cluster
Bielefeld University
E-mail: cimiano@cit-ec.uni-bielefeld.de

Keywords Language resource · linked data · metadata · data harmonization · natural language processing

1 Introduction

The study of language and the development of natural language processing (NLP) applications requires access to language resources (LRs). Recently, several digital repositories that index metadata for LRs have emerged, supporting the discovery and reuse of LRs. One of the most notable of such initiatives is META-SHARE [**?**], an open, integrated, secure and interoperable exchange infrastructure where LRs are documented, uploaded, stored, catalogued, announced, downloaded, exchanged and discussed, aiming to support reuse of LRs[1]. Towards this end, META-SHARE has developed a rich metadata schema that allows aspects of LRs accounting for their whole lifecycle from their production to their usage to be described. The schema has been implemented as an XML Schema Definition (XSD) and descriptions of specific LRs are available as XML documents. Yet, META-SHARE is not the only source for discovering LRs and their descriptions; other sources include the catalogs of agencies dedicated to LRs promotion and distribution, such as ELRA[2] and LDC[3] other infrastructures such as the CLARIN Virtual Language Observatory (VLO) [**?**], the Language Grid [**?**] and Alveo[4], the Open Language Archives Community [**?**, OLAC], catalogs with crowd-sourced metadata, such as the LRE-Map [**?**], and, more recently, repositories coming from various communities (e.g. OpenAire [**?**], EUDAT[5] etc.). The metadata schemes of all these sources vary with respect to their coverage and the set of specific metadata captured. Currently, it is not possible to query all these sources in an integrated and uniform fashion. The Web of Data is a natural scenario for exposing LRs metadata in order to allow their automated discovery, share and reuse by humans or software agents and the benefits of this model including interoperability.

In this paper, we are concerned with how to facilitate the discovery of language resources for a particular task. Given the fact that metadata records for resources are distributed among different catalogs and repositories makes the task of finding a particular resource meeting certain requirements very challenging. This is shown by the fact that many emails to dedicated mailing lists such as the copora mailing list contain requests for resources meeting certain desiderata with respect to type of resource (corpus, dictionary, parallel text, etc.), language, size (in tokens or sentences), etc. So far, no repository that allows discovery of resources across repositories has been available. We have closed this gap by developing Linghub, a linked data based portal that

[1] `http://www.meta-share.eu`

[2] `http://www.elra.info`

[3] `https://www.ldc.upenn.edu/`

[4] `http://alveo.edu.au/`

[5] `http://www.eudat.eu`

indexes and aggregates metadata entries from different repositories including META-SHARE, the CLARIN VLO, LRE-Map and Datahub.io.

Parts of the Linghub as a technical system has been described before [?,?]. Linghub not only indexes the metadata entries, but also harmonizes the information by mapping it to standard semantic web vocabularies as well as to a recently created ontology of language resources[6] that has been developed on the basis of the existing META-SHARE schema [?]. For this, it relies on state-of-the-art word sense disambiguation methods to support the normalization of data.

One of the crucial technologies that enables this integration is that of RDF [?] and linked data. Linked data is based on four fundamental principals [?].

1. Use Uniform Resource Identifiers (URIs) to identify everything in a resource, thus ensuring that every element of the resource can be identified in a standard manner.
2. Furthermore, use HTTP URIs as they require the association with domain names, ensuring that the data can be clearly traced to its host and thus someone responsible for that dataset.
3. Ensure that URIs resolve, in the sense that when typed in to a web browser an appropriate description of the resource is obtained. Ideally the server should detect (using content negotiation) the type of the user and provide HTML results for humans and RDF (serialised in XML or JSON, for instance) for software agents.
4. Provide links to other resources so that it is possible to identify commonalities between resources and to handle issues of semantic interoperability and provenance.

Linked data makes a highly appropriate model for the task of integrating information about language resources as it is natural that this would be handled by means of a web portal and thus stable URIs for resources are easily decided. It is obvious that HTML descriptions should be provided and in order to meet use cases for automatic training of NLP systems, we find the provision of a machine-readable API also of vital importance. Finally, links to other resources are vital to not only provide links back to the source records, but also to ensure that users can find resources for their needs.

In this paper, besides describing Linghub and the semantic normalization methods used, we provide an evaluation of the ability of Linghub to answer the needs of actual users seeking language resources meeting certain criteria. For this, we have analyzed user requests for resources issued on Corpora List and analyzed in how far Linghub is able to answer them. This evaluation is the main contribution of the current paper. We are not aware of any similar evaluation conducted in the context of repositories of language resources, so that to our knowledge this is the first attempt to evaluate the ability of a repository to answer requests for language resources.

[6] `http://purl.org/net/def/metashare`

This paper is structured as follows: Firstly, in Section 2 we will discuss some of the existing related work in particular focussing on the metadata repositories that we will integrate and in Section 3 we will describe how we collected the data. In Section 4 we derive a single data model based on existing standards that will allow us to combine all the resources and then in Section 5 we will show our procedure for harmonizing these resources. We will describe the web portal in Section 6. In Section 7, we provide a thorough evaluation of the system based on real-world queries and thus show the effectiveness of our approach, and finally we conclude in Section 8.[7]

2 Related Work

Harmonization has been a topic of significant importance in managing metadata and it has been noted in several domains that the task of combining metadata from multiple sources. Nilsson [?] proposed a framework for this, where he argues that syntax is of little importance and that harmonization primarily needs to be conducted on a semantic level and he concludes that RDF forms a good model to achieve this. Khoo and Hall [?] worked on merging the Internet Public Library and the Librarians Internet Index and conclude that such resource is very 'ad-hoc' and 'resource intensive'. Nogueras et al. [?] similarly developed 'crosswalks' for geographic data and stress the need for formal modelling such as ontologies for verifying such crosswalks. Chan and Zeng [?] also focus on the use of crosswalks and its use in provding optimal access to data.

One of the key issues in this has between 'monolithic' metadata schemas where one organization has developed and maintained a single and fixed schema for representing metadata. Such standards, frequently defined by fixed XML DTDs and can be useful in providing a minimal level of description between multiple resources however they cannot by their very nature capture all the complexities of the representation of metadata. A highly successful instance of this is Dublin Core [?], where a small fixed set of properties has found use in a wide variety of applications. Instead, some experts have recommended open schemas [?]:

> "A larger set of ontologies sufficient for particular purposes should be used instead of a single highly constrained taxonomy of values."

On the other hand open schemas are difficult as they can often lead to a large number of identifiers for the same properties or concepts, which has lead to the failure of such schemes in the past [?]. The proposal of linked data to allow for additional properties to be dereferenced and thus for extra information about their semantics to be extracted has formed the basis of the Semantic Web. In this context, there have been several vocabularies defined

[7] Some results in this paper have been previously published in the following workshop papers [?,?,?]. The results and methods are expanded and combined in this paper in line with the journal guidelines. In addition, Section 7 provides a novel evaluation of the system.

for representing metadata about data resources, notably the VoID model [?], the DCAT model [?] and its recent extension DataID [?]. These have the possibility to enable more flexible description of language resources and it has even been suggested that these resources could be automatically collected [?].

In the context of language resources, there have been a number of attempts to collect generic metadata about language resource. Firstly META-SHARE [?,?], which has relied on creating descriptions of language resources manually but has still be able to describe in detail a large number of language resources. Another approach that relies on institutional collection of metadata was taken by the CLARIN project, whose *Virtual Language Observatory* [?] collects metadata from a number of host institutes by means of OAI-PMH [?] with a small amount of harmonization provided by the *CMDI Component Specification Language* [?]. A similar project called SHACHI [?] has worked on collecting resources on Asian languges. Another approach in this area is the use of International Standardized Language Resource Numbers [?, ISLRN], where basic metadata has been collected about each resource and they have all been assigned a single number.

As an alternative to the institutional approach, some resources have relied on self-reporting of resource metadata, most notably the LRE-Map [?], which collects information from authors at major research conferences in computational linguistics and as such they are able to collect information on a wide variety of language resources but often leads to quality issues. A similar project, the Open Language Archive Community [?, OLAC] is inbetween both approaches collecting resources from a wide community but trying to bring them into a very fixed schema for their resources. Finally, we note the work of the *Open Linguistics Working Group* [?] a community, which has promoted the use of open data and produced a 'cloud diagram' showing the adoption of linked data language resources over the last four years.

3 Data Collection

In order to realize the goal of providing comprehensive metadata about language resources, it is necessary to collect metadata from a wide range of sources. In particular, we chose four main sources, primarily because these resources have been released under an open license. These resources are:

META-SHARE . A resource created by the META-NET project, the resource provides deep and detailed descriptions of language resources that have primarily been constructed by hand.

CLARIN VLO . The Virtual Language Observatory by the CLARIN project is a collection of resources drawn from a wide variety of insitutes participating in the CLARIN project. In general, the data has been manually curated by the individual contributors and only limited integration has been made between the resources. As such the records are very different in detail and size.

Source	Records	Triples	Triples per Record
META-SHARE	2,442	464,572	190.2
CLARIN	144,570	3,381,736	23.4
Datahub.io	218	10,739	49.3
LRE-Map (LREC 2014)	682	10,650	15.6
LRE-Map (Non-open)	5,030	68,926	13.7
OLAC	217,765	2,613,183	12.0
ELRA Catalogue	1,066	22,580	21.2
LDC Catalogue	714	n/a	n/a

Table 1 The sizes of the resources in terms of number of metadata records and total data size

Datahub.io [8]. This site is an instance of CKAN used primarily to track open and linked data. Most of the data is not of relevance to language resources and as such we apply filters to extact the relevant datasets.

LRE-Map . The LRE-Map was collected by participants at several NLP conferences in the last few years. Unfortunately, only the data from LREC-2014 is available under an open license and the method of collection has lead to significant issues with data quality.

In addition, we have a number of other sources that we investigated for the experiments described in this paper but unfortunately cannot release through the Linghub portal due to licensing:

LRE-Map . Several other conferences of data are accessible on the Web and we have scraped the relevant data.

OLAC . The Open Language Archives Community collects a large amount of data, but clearly states that its own data is not “open”. Fortunately most of the data is also available from CLARIN and other sources.

ELRA/LDA . We also experimented partially with the catalogue of resources provided by the European Language Resource Association (ELDA) and the Linguistic Data Consortium (LDC).

In the following section we describe the format of the resources and the difficulty in consolidating them with our model. The overall size of all the resources is given in Table 1.

3.1 META-SHARE

META-SHARE is provided primarily in a format described in [**?**], which is and XML format and contains over 150 elements and as such is a highly complex XML format. We developed a custom invertible framework called LIXR (pronounced ‘elixir’)[9], which allows us to easily and quickly define the conversion between META-SHARE and RDF. As it was mentioned in Section 1,

[8] `https://datahub.io/`

[9] http://github.com/liderproject/lixr

Component Root Tag	Institutes	Frequency
Song	1 (MI)	155,403
Session	1 (MPI)	128,673
OLAC-DcmiTerms	39	95,370
MODS	1 (Utrecht)	64,632
DcmiTerms	2 (BeG,HI)	46,160
SongScan	1 (MI)	28,448
media-session-profile	1 (Munich)	22,405
SourceScan	1 (MI)	21,256
Source	1 (MI)	16,519
teiHeader	2 (BBAW, Copenhagen)	15,998

Table 2 The relative number of resources in each of the schemas used by CLARIN

the elements defined in the META-SHARE schema were defined as an OWL ontology for interoperability purposes, which is described in [**?**].

3.2 CLARIN

CLARIN is also an XML format and is based on the CMDI metadata infrastructure as defined by [**?**]. This consists of a small shared amount of information and a specific schema, which is normally unique to the data provider, with the exception of Dublin Core metadata which is in two common schemas. The total size and applicability of each of the schemas is given in Table 2 and we have developed export scripts for all of the top 10 formats.

3.3 LRE-Map

The LRE-Map is described in [**?**] and is available partly as RDF, in particular the data from LREC-2014 is available. Unfortunately, the integration was not trivial as there were errors in the RDF [**?**], in particular the use of non-resolving URI schemes that had to be corrected. The older data is also available on the web site and we obtained it by scraping the web site but were advised that this data is not available under any open license.

3.4 Datahub.io

Datahub.io is an instantiation of the CKAN software[10] and as such can easily be accessed through the API and the RDF version of each metadata using the DCAT vocabulary can be accessed. As such the import of this data is quick and can be done at regular intervals.

[10] `http://ckan.org`

3.5 Others

In addition, we looked at three other sources that cannot be included in the public release of Linghub due to the licensing issues. These are OLAC, which uses an XML format and much like CLARIN this format uses different schemas for different data producers. Secondly, there are the catalogues of ELRA and LDC. The ELRA data was made available as an XML file and we wrote a converter for it. The LDC data can be extracted by crawling the website.

4 Modelling

As the basis of the modelling for Linghub we took the DCAT vocabulary [**?**]. The DCAT model is centered around the concept of a *dataset*, which has obvious equivalence to many of the elements in the resources we studied. In addition, DCAT models distributions, i.e., downloads, and catalogues and we imported each of these elements. We found that some distinctions made in DCAT, most notably the distinction between access URLs and download URLs, that give the link to the dataset's home page and the direct link to the data respectively were not clear in any of our sources and this will continue to be a major stumbling block to providing fully automatic access to datasets.

The DCAT model, however, only provides for generic descriptions of datasets and we wished to capture specific elements that would be of interest to linguists. As such we worked on developing an extension of DCAT based on the META-SHARE model, which we call the META-SHARE ontology. This resource is described in [**?**], and for the benefit of readers we briefly recap the model here.

DCAT consists of a *catalog* composed of *datasets*, with a *catalog record*, which corresponds to the META-SHARE metadata info element. META-SHARE contains a much richer description of many aspects than DCAT including contact details, version information, validation and proposed and actual usage of the dataset. These elements, when available, were directly added to the model. In many cases, basic properties in the META-SHARE ontology, such as the language of a resource, were to be found nested under several layers of tags and in such cases we added property chain links so that they would be more compatible with other resources. For example, the rights statement of a resource could be found only under the headings "Distribution Info", "Licence Info" and "Licence", and so we added this property also directly to the root data element to comply with DCAT. In addition, META-SHARE contains significant resource type specific information that is defined by the type of the language resource: one of *corpus*, *tool/service*, *language description* or *lexical conceptual resource*. These extra elements include media type (text, audio, video or image) and the encoding of information, formats, classifications, and so forth.

In addition to aligning the META-SHARE model to DCAT and Dublin Core, the META-SHARE ontology improved on the original model in the following ways:

- Removal of the `Info` suffix from the names of wrapping elements of components.
- Improvement of names that created confusion, as already noted by the META-SHARE group and/or the LD4LT group[11]; thus, `resourceInfo` was renamed `LanguageResource`, `restrictionsOfUse` became `conditionsOfUse`.
- Generalization of concepts, e.g. `notAvailableThroughMetashare` with `availableThroughOtherDistributor`;
- Development of novel classes based on existing values, for example: `Corpus` ≡ ∃`resourceType.corpus`
- Grouping similar elements under novel superclasses, e.g. `annotationType` and `genre` values are structured in classes and subclasses better reflecting the relation between them. Indicatively, the superclass `SemanticAnnotation` can be used to bring together semantic annotation types, such as semantic roles, named entities, polarity, and semantic relations.
- Extension of existing classes with new values and new properties (e.g. `licenseCategory` for licences).

5 Harmonization

Due to the variety of source from which we are obtaining metadata it is inevitable that there are differences between them. Moreover, the quality of the resources varies greatly, for example in the case of language META-SHARE uses ISO 639-3[12], but a crowd-sourced resource such as LRE-Map has a wide variety of representations in free text. Our approach focuses on the properties that are most important for using a resource including whether the resource resolves, what license it is available under, the type of the resource (e.g., corpus) and the language or languages covered by the resources (properties may of course have multiple values).

5.1 Availability

Certainly the biggest barrier to re-using a resource is obtaining it and it is unfortunately the case that many resources become unavailable due to server failure or similar reasons. We also note that there is an important distinction that must be made between ‘access URLs’, which is generally a page containing information and documentation about the resource and generally a download link, and the ‘download URL’, where the resource can be directly accessed. If we wish to enable use cases where software agents can autonomously access

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

[12] http://www-01.sil.org/iso639-3/

Format	Resources	Percentage
HTML	67,419	66.2%
RDF/XML	9,940	9.8%
JPEG Image	6,599	6.5%
XML (application)	5,626	5.6%
Plain Text	4,251	4.2%
PDF	3,641	3.6%
XML (text)	3,212	3.2%
Zip Archive	801	0.8%
PNG Image	207	0.2%
gzip Archive	181	0.2%

Table 3 The distribution of the 10 most used formats within the analyzed sample of URLs. Note XML is associated with two MIME types.

resources we would need the latter type of URL, however sadly at the moment nearly all links given in our sources are ‘access URLs’ and thus we only analyze these links at the moment.

In our study, we then accessed 119,920 URLs given among our sources of metadata and we found that 95% of these resolved successfully (i.e., HTTP Response was 200 OK). We then also analyzed the reported content type of the response, the results of which are in Table 3. We found that text formats, in particular HTML tended to be the predominate format and we would assume that these correspond to human-readable pages and not the actual resource in the most case. A little of 14% of resources are in a format that seem like data sources although we observed that many of the RDF/XML results were from source providing Semantic Web compatible descriptions of resources. In addition, some number of images were found, which were generally scans of historical documents.

5.2 Rights

While accessing the resource itself is one of the main goals of any user of language resources, any responsible user must take into account the license that a resource is released under. Thus, It is not an infrequent case that users searching LRs want to determine which licenses are acceptable for their purposes. Each of the different platforms providing access to LRs have different means to select the desired licenses.

- The META-SHARE portal offers a faceted browsing where one of the facets is the license declared for the resource. The browsing experience is enhanced by other facets that permit discriminating resources based on their availability (restricted/unrestricted) or by their restrictions of use (like ‘commercial use allowed’). This is a rather satisfactory organization of the information more complete than the other platforms.

- The CLARIN Virtual Language Observatory[13] also offers a faceted browsing, and one of the 8 facets is devoted to the ‘availability’. Many of the resources fall under diffuse categories (such as “open” or “free”) without referring the actual licenses. The metadata describing the license is a free text instead a URI determining the license in use.
- The OLAC Language Resource Catalog offers the search functionality from their portal[14] and a faceted search as well. However, catalogued resources seem to lack this information, existing only three types of licenses (‘CC-BY-ND’, ‘CC-BY-SA’ and ‘others’) making the navigation impracticable. An additional facet for ‘other rights’ performs no better due to the opposite reason: there are so many types of ‘rights’ that it is not possible to search with a given criterium.
- The LREMap resource portal[15] does not permit searching by license. Further, once obtained the resources they include a ‘license’ description but in a textual manner.
- Datahub.io permits selecting the license in the faceted browsing they offer. Although nothing prevents dataset creators from declaring their own licenses, they are driven by the user interface to use one of the pre-determined license-types. This greatly reduces the license proliferation and searching by rights in Datahub is still relatively acceptable.

The best description of the rights information is given by licenses with a well defined URI. If this were regularly the case, the *license profileration* problem would be easily solvable. From this perspective, the META-SHARE portal is the best one; with Linghub having equivalent capabilities only improved for true experts capable of formulating SPARQL queries.

5.3 Usage

The usage of a language resource is an indication of what purpose it was created for. Following the example of META-SHARE we distinguish between intended use and actual use, where intended use is the use by the creator of the resource and the actual use is another application that has used this resource. As the data has very little information on the latter case, we focussed primarily on the intended use, which is recorded clearly in two resources: META-SHARE and LRE-Map. The taxonomies used in each scheme is different with META-SHARE defining 83 possible values and LRE-Map suggesting 28 values, while actually 3,985 values have been used. This is due to the collection method of LRE-Map, which has a dropdown of options or the user can select ‘other’ and enter their own value.

For the 28 suggested LRE-Map values we added a manual mapping to the META-SHARE values and for the rest of the values we developed a mapping

[13] `https://vlo.clarin.eu`

[14] `http://search.language-archives.org/`

[15] `http://www.resourcebook.eu/`

Resource	Label Accuracy	Instance Accuracy
SIL *dice coefficient*	81%	99.50%
SIL *levenshtein*	72%	99.42%
BabelNet *dice coefficient*	**91%**	99.87%
BabelNet *levenshtein*	**89%**	99.85%
SIL + BabelNet		
dice coefficient	**91%**	99.87%
levenshtein	**89%**	99.85%

Table 4 Accuracy of language mappings

algorithm based on using the Snowball stemmer [?] and string inclusion match to detect variants. From a random sample of 100 of such terms we found that 66% were correct matches, 16% were empty fields or non-specific terms (e.g., 'various uses') and 16% were overly general (e.g., 'acquisition'). In addition, we had one false negative (due to a typo 'taggin pos' [sic]) and one novel usage that was not in META-SHARE ('semantic system evaluation'). As such, we conclude that the system has 98-99% accuracy.

5.4 Language

We decided to normalize the language identifiers around the ISO 639-3 standard due to its wide adoption and coverage of nearly all human languages. Many resources used either this standard already or the shorter two-letter codes from ISO 639-1 and as such the primary challenge is in fact mapping the names given in English text. To achieve this we collected lists of language names from the official SIL database[16] and from BabelNet [?], a large multilingual lexicon.

We compared the results using two string similarity metrics, namely the Dice Co-Efficient and the Levenshtein Distance, the results of which are reported in Table 4. We measured the accuracy by constructing a sample of 100 labels and manually mapping them to language codes and we present the results both as total number of labels matched and weighted by the usage of those labels by resources. For both resources, we see very high accuracy and the unmapped labels were mostly for those labels that were very rarely used.

When deploying this system in the Linghub, however, we did notice that the system made one very noticable error, namely mapping the label 'Greek' to the language (Muscogee) 'Creek' as we only had labels for the language 'Modern Greek'. We thus applied a second scan adding this and a few other common language name variations.

[16] `http://www-01.sil.org/iso639-3/download.asp`

Resource	Duplicate Titles	Duplicate URLs
CLARIN (same contributing institute)	50,589	20
Datahub.io	0	55
META-SHARE	63	967
LRE-Map	763	454

Table 5 The number of intra-repository duplicate labels and URLs for resources

Resource	Resource	Duplicate Titles	Duplicate URLs	Both
CLARIN	CLARIN (other contributing institute)	1,202	2,884	0
CLARIN	Datahub.io	1	0	0
CLARIN	LRE-Map	72	64	0
CLARIN	META-SHARE	1,204	1,228	28
Datahub.io	LRE-Map	59	5	0
Datahub.io	META-SHARE	3	0	0
LRE-Map	META-SHARE	91	51	0
All	All	2,632	4,232	28

Table 6 Number of duplicate inter-repository records by type

Duplication	Correct	Unclear	Incorrect
Titles	86	6	8
URLs	95	2	3
Both	99	1	0

Table 7 Precision of matching strategies from a sample of 100

5.5 Type

By the type of resource we mean the form of the resource, such as the basic categorization of META-SHARE into 'Corpus', 'Lexical Conceptual Resource', 'Lexical Description' and 'Tool/Service'. For this, we used the properties from existing resources and applied the Babelfy linking algorithm [**?**]. Once we had this result, we manually selected those senses that corresponded to language resources, giving us in total 143 categories of which the top 10 were: Sound', 'Corpus', 'Lexicon', 'Tool' (software), 'Instrumental Music'[17], 'Service', 'Ontology', 'Evaluation', 'Terminology' and 'Translation software'.

5.6 Duplicate detection

It is a natural effect of collecting records about resources that we will have multiple records that in fact describe the same resource. In order to provide

[17] These resources are in fact recordings of singing in under-resourced languages

a single view of the description of a dataset for a user, it is important that all the information can be consolidated so that we can for example, indicate all uses of a particular dataset. This is particularly the case with LRE-Map where resources are frequently reported multiple times in different uses. As such, we make a fundamental distinction between *inter-repository duplication*, where we have two records from different sources describing the same entity, and *intra-repository duplication*, where resource are described multiple times by the same source. Note in the case of CLARIN it is quite common to see duplication between different contributing instances and these are normally descriptions of the same resource in different formats, so we treat them as inter-repository duplicates. In both cases, we based our detection on looking for duplicate title and URLs for resources.

We will first look at the case of intra-repository duplication, where the causes seem to be quite different in each of the sources:

META-SHARE The duplicates here were due to errors in the export and were easy to correct.

CLARIN In many cases sequences of resources had multiple records. For example the 'Universal Declaration of Human Rights' had an individual page for each language and thus we merged resources with the same title, as we believe this is of more use to our users.

Datahub.io This resource does not allow for duplicate titles, but duplicate URLs are quite common, however these are more likely due to shared resources (e.g., several resources use the same SPARQL endpoint).

LRE-Map Duplicates in LRE-Map are caused by multiple submissions using the same resource, and as such we wish to aggregate all these citations in order to make it clear how frequently a resource is used and thus show the resources' quality

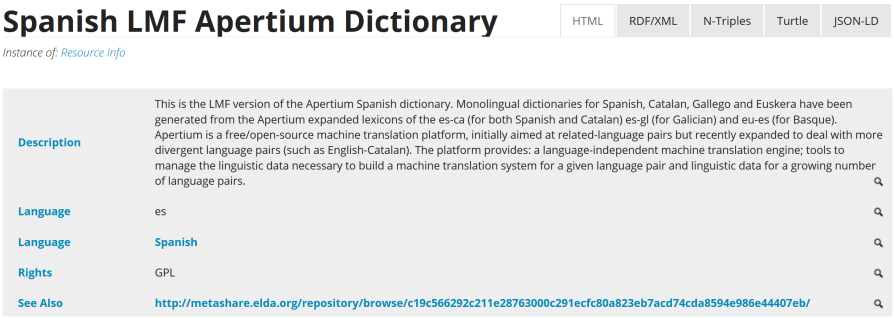
The total number of intra-repository duplicates detected is presented in Table 5.

For the case of inter-repository duplication, we assume that in all cases this is due to multiple descriptions of the same resource. In order to evaluate the effectiveness of our method of matching by title and URL we took a sample of 100 resources and examined whether they actually referred to the same resources. The results of this analysis are given in Table 7 and the number of duplicates detected in total is given in Table 6.

6 Linked Data Interface

Proper access to the dataset requires a pleasant and functional interface and the design of such an interface should not be ignored. As such, we provide Linghub as a website (a screen shot is shown in Figure 1)[18]. that allows both humans to see the data in the form of HTML pages and for computers to

[18] `http://linghub.org/`



Spanish LMF Apertium Dictionary

Instance of: Resource Info

Description	This is the LMF version of the Apertium Spanish dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan). The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.
Language	es
Language	Spanish
Rights	GPL
See Also	http://metashare.elda.org/repository/browse/c19c566292c211e28763000c291ecfc80a823eb7acd74cda8594e986e44407eb/

Fig. 1 A screenshot of the Linghub interface

access that data in the following formats: RDF/XML, Turtle, N-Triple and JSON-LD. Simple templates are used that render the data in HTML so that human users can obtain a consistent view of the data, from all of the sources displayed in Linghub. Furthermore, we provided some extra mechanisms, by which users could access the data

Faceted browsing . Users can choose to slice datasets up by a number of elements including language, rights, type, creator, source, contributor and subject. The users can see all relevant dataset as well as querying specific parameters

Free-text search . The most common search method employed on the Web is free-text search and we employ this by building a search interface that indexes all the literal values in the data and allows free search over them. In addition, for certain values, especially languages we added in the values of the literals and indexed them.

SPARQL search . Finally, we enabled SPARQL search for advanced and automated users, that as we show in the evaluation improves the ability of clients to find relevant results. For performance reasons, we limit the expressiveness of queries to those that are likely to be easy to compute. The endpoint by default returns results in JSON [**?**].

7 Evaluation

In this section we describe an experiment that measures the ability of Linghub to solve real users needs on the basis of a set of real requests of language resources. Both the traditional (free text) and advanced (SPARQL) search capabilities of Linghub have been considered.

7.1 Evaluation Methodology

To evaluate the usefulness of Linghub based on actual needs for linguistic resources, all questions for language resources posed on the Corpora Mailing

List (CML) [19] from January 1st 2015 till June 3rd 2015 were examined. After ruling out a small subset of questions on the mailing list that were too underspecified and unclear to be operationalised even for a completely manual search, a catalogue of 23 request was assembled. Each request poses a number of constraints on the data being searched. Constraints typically pertained to the *type* of the resource like corpus, lexicon, tool or service; the *language* the resource is applicable for; the *extent* of the resource measured in various metrics like word count and additional qualities, e.g. intended use.

Each request was then searched for in Linghub using both the standard searchinterface as well as SPARQL queries. This covers both relevant audiences: Standard users like most linguists and developers looking for data without being familiar with SPARQL, and Semantic Web experts fluent in SPARQL. Results of the queries were counted and each was evaluated as relevant, irrelevant or related[20] to the query. The original request #6 was skipped, because the evaluators could not agree what exactly constitutes a “corpus suitable for training hierarchical classification models”.

7.1.1 Limitations

The methodology chosen can only evaluate Linghub in a limited way. The first and most important limitation is the biased sample of queries. People asking for help on the Corpora Mailing List will most probably already have searched using standard means, such as Google or repositories known to them. This can limit the queries to very specific requests, that are complicated to find answers for. The technical expertise of the users and, consequently, their more specific requirements also contribute to the requests being rather intricate. Thus, all but a few queries examined are expert level queries, limiting the general applicability of this evaluation. Although the timespan examined constitutes a reasonable sample with six months of mailing list requests analysed, the number of queries is quite low. However, as section 7.2 will detail, their overall type and language profile matches available Linghub data. Due to the complicated nature of the requests, translating them from natural language to concise and reasonably narrow queries to the search interface respectively SPARQL constitutes another hurdle, that is hard to control for. Search was furthermore conducted by two persons, whose interpretation of the meaning of the query and thus the relevance of the results is a subjective judgement made by agreement of the two annotators.

7.2 Resource request analysis

The requests were analysed in regard to the constraints they express. All queries expressed at least 2 constraints, usually regarding the resource type

[19] `http://mailman.uib.no/public/corpora/`

[20] “related” meaning that some constraints were not or not fully met.

Language count constraint	CML requests	LH resources
1	8	51350
2	1	884
3	2	69
4	1	31
$\geq 5, \leq 19$	0	54
> 20	2	4
unspecified	7	635895

Table 8 Language constraints in requests and resources available in Linghub covering specific numbers of languages

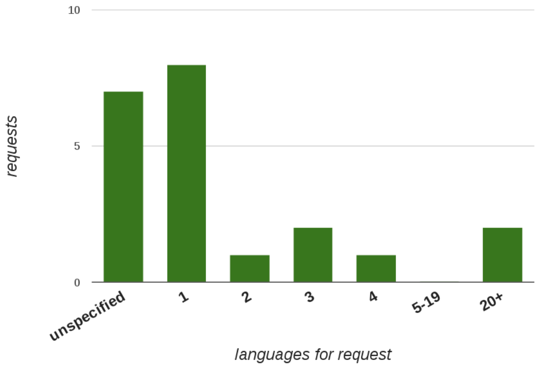


Fig. 2 Distribution of number of languages in requests to the Corpora Mailing List

and the preferred language of the resource. Requested resource types were limited to corpora, tools, lexical and spoken resources. Corpora were requested in nearly 70% of the cases, followed by tools (17.4%) and lexical resources (13.0%). As seen in Figure 2 and 3 and Table 8, this resembles the resource distribution in Linghub, where 71.5% are corpora as well. Most language restrictions asked for one language or made no specification at all. The decreasing number of requests for higher language restriction count with a small rise above 20 languages again resembles the resource distribution in Linghub, that also shows a power law distribution of resource language count. The high number of resources for 20 or more languages can be explained by the use of European corpora that contain all European languages.

Besides these common constraints, most requests also define at least one further restriction. For corpus requests, a common restriction is having a large size in number of words. Specifically, gigaword corpora or corpora with billions of words are highly searched for. Surprisingly, only one request explicitly limited the license to free and open resources. Further restrictions mention the type of annotation (for example manually checked or annotating specific features), the feature set of a tool or service or the semantic content of the resource. These restrictions can only be judged by first doing a full-text search in the description of the resources, then reading these descriptions and judg-

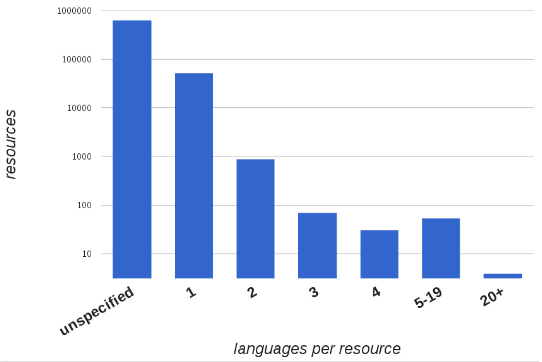


Fig. 3 Distribution of number of languages in Linghub ($\log_{10}$-scaled)

ing their appropriateness. Thus, the assessment of relevance or relatedness of resources found during search is subjective and may not always be comparable.

7.3 Standard search interface

While trying to express the facets of the questions from the CML in Linghub, several shortcomings both of the browsing functionality and the 'simple' search became apparent. The inability to declare combined restrictions for several facets (description, language, rights, . . .) simultaneously is presumably the most crucial limitation for both the browsing interface and the search form. Online catalogues usually either offer alternative search form for 'advanced' users where a flexible number of pairs of field specifiers and a corresponding search patterns can be defined, whose constraints will be combined by boolean operators or complex search patterns can be defined, allowing for sub-patterns that are matched against solely against specified record fields, e.g.:

```
(title:corpus OR description:corpus) AND
  (description:(part of speech OR pos)) AND
  (language:(deu* OR german))
```

Since the overwhelming majority resource requests from the CML involved soft constraints that can only be expressed with full-text patterns against resource descriptions, additional structured information about language or rights could not be harnessed during the query (as the sole facet choice already had to be used for the description). Especially for formulating multiple soft constraints against the resource descriptions the full text search capabilities from SQLite exposed in the search interface proved very useful.

Some of the facets offered by Linghub are naturally textual (e.g., title, description, creator), but others are rather categorial with a limited set of options (e.g., resource type, language). The search interface still has some limitation that were discovered by the annotators including the lack of an ability to set constraints on multiple property facets simultaneously, the fact

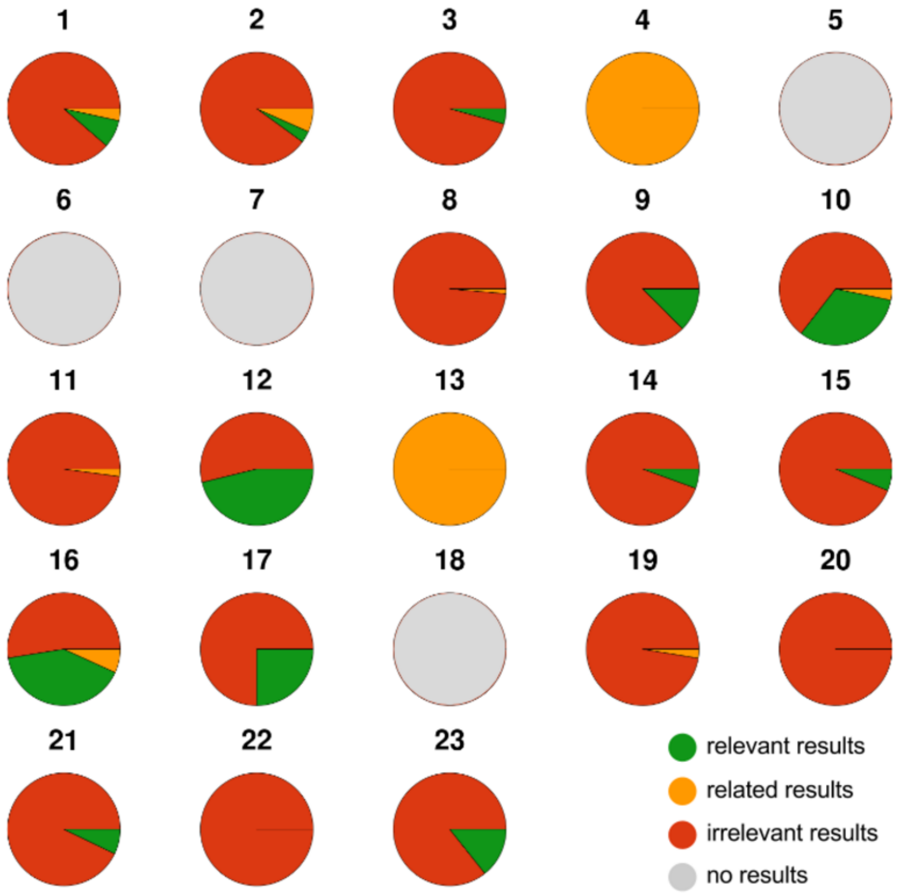


Fig. 4 Percentages of relevant standard interface search results

that categorial facets must be queried as full-text, without any knowledge of the values and that free-text search also indexes descriptions in languages other than the interface language of Linghub[21]. These criticisms will be addressed in future versions of the interface.

Excluding the queries that did not yield results, on average 9.7% of the results were relevant, and further 2.6% were related. The high share of irrelevant search results does not come surprising, as the queries formulated for evaluation generally were rather open than restrictive in many cases to favour recall over precision, as one can assume that potential users of language resources will be willing to invest a bit more time to manually sift through an acceptable amount of additional false positives rather than risk missing information of an additional potentially useful resource. The results are shown in Figure 4 and Table [**?**].

[21] Currently, the language interface of Linghub is English

7.4 SPARQL search

While the standard search interface presents the default access point to the Linghub knowledge base, SPARQL search is regarded as "advanced search". Using SPARQL naturally solves most issues of the standard search interface, as it can be used for granular search in arbitrary literal fields and natively provides logical operators for filters to combine them. If values are modelled as classes, the user does not have to use string matching but can use object relations, working on a well-defined, semantic level. The only disadvantage of SPARQL is its inaccessibility to most linguists that don't have knowledge of database query languages.

SPARQL queries were written and executed using the downloaded Linghub dump for convenience. The queries and detailed results can be found in the appendix.

Figure 4 shows the results for each of the queries. Excluding queries 6 and 18 that yielded no results, on average 22.0% of the results were relevant, 8.7% were related. Roughly a quarter of the queries failed, yielding no relevant results or no results at all. In general, this can be considered as a good result, taking into account the complicated nature of the queries. Specifically, it indicates that the more granulated means SPARQL provides to the user lead to more accurate results. The results are presented in Figure 5 and Table 9. The SPARQL results are both more accurate and also find more results in total, finding 79 results as opposed to 62 for free-text and a further 20 related results with only 17 further results for free-text.

Language proved to be the most restricting constraint, leading to most irrelvant results, even if other constraints are met. Although SPARQL was used to granularly restrict the languages by leveraging the `dc:language`, queries were not limited to this relation but also incorporated searching for the language name in the `dcterms:description` to increase recall. Size was another important constraint that was never fully met. If size was a restriction, it always was in the range of billions of words, excluding all relevant corpora as too small. Corpora with millions of words were counted as related. Again, description texts were used to retrieve hints on corpus size. One large advantage of SPARQL was being able to explicitly address values for several facets simultaneously, such as distribution and license information, allowing to filter for free and open resources, as well as links to the data itself. The selection of the queries in CML reduced this use-case to one case, but it seems to be an important and often overlooked facet of data acquisition.

7.5 Data Completeness and Quality

To obtain statistics on data completeness, relative frequencies of Linghub resources carrying at least one property-value pair for the various facets offered by the search frontend were determined. The basic quantity of Linghub resources was defined as all URIs in the Linghub dataset with the linghub.lider-

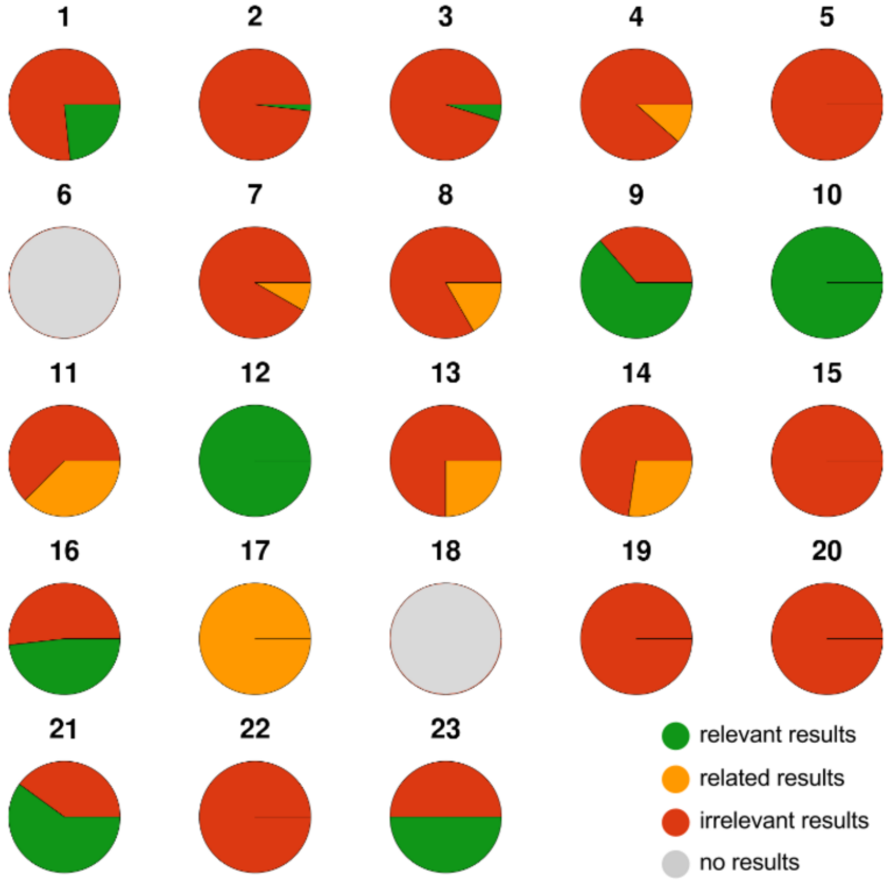


Fig. 5 Percentages of relevant SPARQL search results

project.eu hostname[22] appearing in subject position of at least one triple. Figure 6 and Table 10 presents the aforementioned relative frequencies.

These statistics reveal significant variation in coverage over the various facets, that can result in unexpected recall when relying on facets with low coverage. To illustrate this with an example: 444 Linghub resources containing keywords 'spanish' or 'spain' in their description also carry a corresponding dc:language property. On the other hand 493 resources with aforementioned keywords in the description do not carry a dc:language attribute. Although the mere appearance of the keywords are not conclusively indicative that the resource should be counted to the corresponding language, the majority of the latter resources appeared to be Spanish or relevant for Spanish when examining a 10% sample. Combined usage of the language property values when present

[22] At the time of the evaluation Linghub was only available at `http://linghub.lider-project.eu/`. The preferred URL is now `http://linghub.org`

Query	Search Results Count	Percentage Relevant	SPARQL Results Count	Percentage Relevant
1	88	11.36%	60	23.33%
2	60	10.00%	63	3.17%
3	23	4.35%	21	4.76%
4	1	100.00%	4	100.00%
5	0	0.00%	16	0.00%
7	0	0.00%	6	16.67%
8	73	1.37%	24	16.67%
9	28	14.29%	11	63.64%
10	31	35.48%	20	100.00%
11	47	2.13%	16	37.50%
12	13	46.15%	4	100.00%
13	1	100.00%	4	25.00%
14	18	5.56%	11	27.27%
15	16	6.25%	2	0.00%
16	59	47.46%	56	48.21%
17	4	25.00%	1	100.00%
18	0	0.00%	0	0.00%
19	39	2.56%	20	0.00%
20	81	0.00%	4	0.00%
21	42	7.14%	5	60.00%
22	1	0.00%	9	0.00%
23	14	14.29%	1	100.00%
Average	27.78	18.84%	15.57	35.92%

Table 9 Search results by relevance for Free-text and SPARQL search

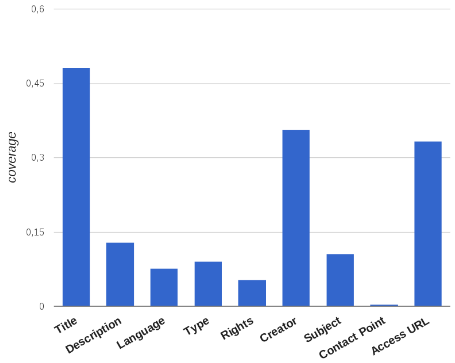


Fig. 6 Portions of Linghub resources with at least one property value for the main facets

and fall-back to text-matching on title and description, as used in the SPARQL queries, can mitigate this problem. Using the free text pattern syntax with field specifiers, this procedure can be sketched as:

```
language:spa OR title:(spanish OR spain*) OR
  description:(spanish OR spain*)
```

Required Facet	Absolute Freq	Relative Frequency
(none)	688287	100%
Title	331199	48.12%
Description	89053	12.94%
Language	52392	7.61%
Type	62063	9.02%
Rights	36869	5.36%
Creator	244725	35.56%
Subject	72768	10.57%
Contact Point	2436	0.35%
Access URL	229020	33.27%

Table 10 Portions of Linghub resources carrying at least one property value for the respective required facet

Analogous text-pattern fallback strategies might be employed for other properties with low-coverage (e.g., type) and could be offered as on-demand option in the search interfaces.

Several of the examined resource requests from the CML asking for corpora also formulated minimal requirements for their size. Hence adding size descripction as another explicit, structured search facet for Linghub would appear beneficial. About 4000 resources listed in Linghub carry property-value pairs for numeric values quantifying their size according to a specified unit. Excluding also cases where the unit is not clearly specified or where the size value is just a sentinel value for no available, about 2490 resources with well defined, structured size information remain (all of these originate from META-SHARE). Increasing this coverage towards a substantial portion of corpus resources indexed in Linghub would be quite beneficial. However, although missing size information could probably be extracted from description texts for many resources, achieving a satisfactory level of correctness of such a process would presumably require a prohibitively great extend of manual annotation and curation.

8 Conclusion

In this paper we have presented a consolidation of a large amount of information about language resources from various source of differing levels of quality and granularity of information. We found that, while some properties are generally quite feasible to reconcile, in many cases the information contained in the metadata is quite insufficient, especially in the case of licenses of resources. We then developed an interface to this reconciled dataset using linked data principles and we evaluated this relative to real user queries made on the Corpora Mailing List. We saw that the SPARQL querying allowed more accurate and complete querying but this interface may prove difficult for many working linguists and an open question remains of how to provide these benefits to all users of Linghub.