

Linghub2: Language Resource Discovery Tool for Language Technologies

Cécile Robin*, Gautham Vadakkekara Suresh*, Víctor Rodríguez Doncel†,
John McCrae*, Paul Buitelaar*

*Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway
Lower Dangan, Galway, Ireland

†Polytechnic University of Madrid

Avenida Ramiro de Maeztu, s/n, Madrid, Spain

{cecile.robin, gautham.suresh, john.mccrae}@insight-centre.org
paul.buitelaar@nuigalway.ie, victor.rodriguez@upm.es

Abstract

Language resources are an essential component of natural language processing, as well as related research and applications. Users of language resources have different needs in terms of format, language, topics, etc. for the data they need to use. Linghub (McCrae and Cimiano, 2015) was first developed for this purpose, using the capabilities of linked data to represent metadata, and tackling the heterogeneous metadata issue. Linghub is aimed at helping language resources and technology users to easily find and retrieve relevant data, and identify important information on access, topics, etc. This work describes a rejuvenation and modernisation of the 2015 platform into using a popular open source data management system, DSpace, as foundation. The new platform, Linghub2, contains updated and extended resources, more languages, and continues the work towards the homogenisation of metadata through conversions, through linkage to standardisation strategies and community groups, such as the Open Digital Rights Language (ODRL) community group.

Keywords: linked data, data repository, language resources, linguistics, open digital repository

1. Introduction

Discovering and searching for adequate Language Resources (LRs) (including services) is at the core of any Natural Language Processing (NLP) task. The importance of these steps should not be underestimated, as they have a significant impact on the technologies built from them. Several portals exist, but do not provide solutions that takes into account data quality, data coverage and data standardisation all at once. The omission of any of these aspects reduces the chance of the data being discovered and used by stakeholders.

The Linghub (McCrae and Cimiano, 2015) platform was first developed in 2015 in response to a lack of the aforementioned aspects. It makes use of the capabilities of Linked open Data (LOD) to represent metadata, using standards so that users does not need to learn new formats. Linghub’s main objectives are to provide information on data that is open access, traceable, that can be queried using the LD query language SPARQL, and that tackles the issue of heterogeneous metadata. In this new version of Linghub, Linghub2¹, we push this further, by creating an improved user experience with advanced and stable search capabilities either through the Web interface or SPARQL queries, by integrating the use of standards and of quality tests, and by providing more data with previous repositories and new datasets all available under the same interface and described using the same schemas.

The work described in this paper is a result of the H2020 Prêt-à-LLoD project² - Ready-to-use Multilin-

gual Linked Language Data for Knowledge Services across Sectors (PAL). This project aims at providing standards and ready-to-use multilingual tools and resources for the development of Natural Language Processing (NLP) tasks and technologies. In Section 2 we will give an update on the related work in language data portals, then Section 3 will describe the main objectives of the new platform. Next, in Section 4, we will explore the different functionalities, in terms of search and query of the data, import functions, storage or formats, which tackle some of the objectives. Furthermore, in Section 5 we will present some test and conversion components built to improve quality, standardisation, integration, and ease of use of the resources from various sources. Finally, in Section 7 we will expose some of the challenges faced during the development, and some expectations that could not be met.

2. Related Work

In terms of other related data portals, little has changed since 2015 in the offering of the portals described in (McCrae and Cimiano, 2015), and the way they operate. Only Datahub³ has changed its strategy and now only provides metadata for which they can offer the datasets, which significantly reduces their offer. They also put in place a premium access which gives extra benefits. However we strongly believe in the open access of the portal information, in order to reach as many users as possible, and therefore help further development and extensions of the resources by the community.

¹<https://new.linghub.org/>

²<https://pret-a-llod.github.io/>

³<https://datahub.io/>

A significant initiative has emerged recently from collaboration within the Language Technology (LT) community in Europe (including industry, innovation and research): the European Language Grid (Rehm et al., 2020) (ELG). The ELG initiative⁴ is interested in the growth the community, and a centralization of European LT activities in order to foster collaborations. As part of these objectives, ELG has developed a LT platform⁵ that is using a grid architecture. ELG is not based on LOD, but has instead developed its own metadata format, ELG-SHARE. Contributors of a resource are required to familiarize themselves with the custom built metadata schema, and then to manually convert their data using the schema in an XML file. This entails a significant amount of work on the part of the user. This also means that the schema chosen is not universal and is only known to collaborators of ELG. By following LD standards in Linghub, we facilitate interoperability.

3. Linghub2 Objectives

Linghub2 is an outcome of the PAL project. One of the project objectives is to solve the issues of data cleaning, organizing and collecting, which can take up to 80% of the time needed to develop a language technology. The consortium believes some of it can be alleviated by allowing the discovery of LRs across multiple repositories through a single platform that meets a number of criteria: usage of expressive and reusable metadata for better integration (ie. based on Linguistic Linked Open Data (LLOD)), providing solutions to the heterogeneity of metadata acquired through various portals, allowing efficient search and discovery of language data, clear information about right of use, availability, technical quality, etc.

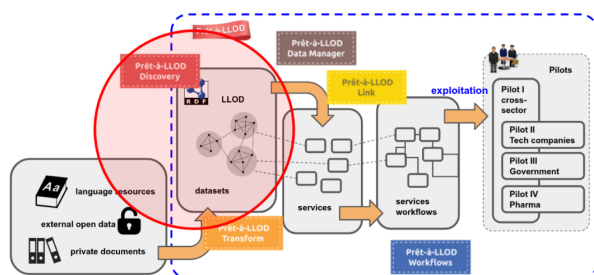


Figure 1: Prêt-à-LLOD project pipeline

For the provision of this new version of Linghub, a refactoring of the previous platform was selected so that it uses state-of-the-art solutions for creating open access repositories. Stability and sustainability, as well as pre-built tools for data management and search also motivated this choice. The new platform aims to extend the capabilities of the previous version, more user-friendly and robust and with extended features. Just

⁴<https://www.european-language-grid.eu/>

⁵<https://live.european-language-grid.eu/catalogue/>

like Linghub version 1, Linghub2 is a Web-based platform that is open access, sustainable and provides LRs and services from various repositories. In addition, this version integrates quality tests and conversions to standards, and offers many ways for the user to explore and search the data using the discover feature, the multilingual search facility or SPARQL queries.

4. Linghub2 Functionalities

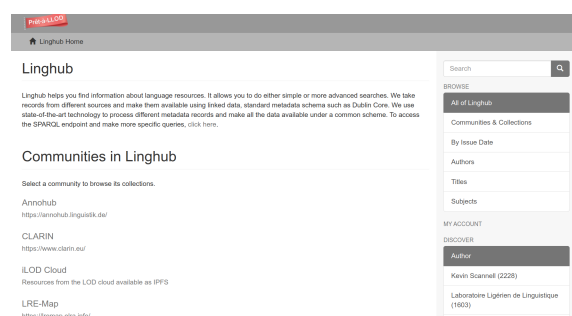


Figure 2: Linghub2 Homepage

4.1. DSpace-Based Platform

We chose DSpace⁶ as the data management software to create the new Linghub2 platform, and on which we customised the configurations and provided extensions. DSpace is open-source, which was one requirement, and widely used by academic, non-profit and commercial organisations. It is an easier solution to install and customize compared to other open source software for Digital Libraries (Khan, 2019). It is maintained, developed and supported by a strong user community, with the help and guidance of DuraSpace⁷ and provides various forums to connect with the community.

DSpace is designed to allow an easy and open access to all types of digital content, although in our case we are only interested in the metadata of datasets.

The metadata stored in DSpace is LD-based, which was a requirement for Linghub2, to ensure standardisation and integration of distinct sources within the same schema. DSpace is based on the standard Dublin Core (DC) schema to represent the core metadata of LRs, although it is possible to add other schemas and fields.

4.2. Customized UI

The standard DSpace template was customized to Linghub and Prêt-à-LLOD themes as can be seen in Figure 3. Several modifications were made to the UI to display the resources as described in the following sections.

⁶<https://duraspace.org/dspace/>

⁷<https://duraspace.org/>

4.3. Search / Browsing / Filtering

Thanks to Apache Solr⁸, DSpace provides a built-in search facility that allows the user to easily look for specific content within the metadata. There is a free-text search which fetches the appropriate content within any metadata field, and the advanced search which allows the user to make specific queries, using filters on a set of metadata (eg. language, license). Figure 3 shows an example of a search for a corpus of spoken conversations in Zulu, using the free text search and filtering on *Type* and *Language(ISO)*. In addition to the default DSpace filters, we added the ISO 639 language codes⁹ and licenses in the list, as they are essential components for the search of LRs.

It is also possible for the user to simply browse and discover data using a sidebar displaying a selection of metadata available in the resources.

4.4. Import

DSpace has different options for importing large amounts of data and metadata, and for editing batches of metadata in the platform. Even though several methods are available to import data, the data is stored as LD. We chose the CSV import method for its simplicity of format. Since we are planning to update the platform with additional resources over time, we are aiming to simplify the procedure of conversion as much as possible, so that it can be adopted by a wide range of users. The LD-based metadata information is given as column titles (eg. *dc.title*), and each row contains the values for one resource (eg. *Universal Declaration of Human Rights*). DSpace creates a unique identifier (*dc.identifier.uri*) for each resource, which will be used as subject uri for the RDF triples.

4.5. SPARQL Endpoint

As LD built-in capabilities, DSpace supports the publishing of stored content as Linked Open Data (LOD) in a triple store of choice. A triple store comes with a SPARQL¹⁰ endpoint, which allows the user to make formal SPARQL queries on the data. The well established SPARQL server Jena Apache Fuseki¹¹ (Fuseki) is recommended and pre-configured to be used in DSpace, which we therefore chose as solution for our triple store. The SPARQL endpoint available in Fuseki that we integrated in DSpace allows human users or machine-based agents to query Linghub2 and retrieve large quantities of data. (see Figure 5).

This offers yet another search facility that benefits from RDF.

4.6. Sustainability

The maintenance and sustainability of the platform is ensured by a combination of documentations and pro-

cedures. A detailed technical documentation on how to set up the platform, import and maintain the data was created. Version control of the platform is provided through a Github¹² repository, part of the general Prêt-à-LLOD project repository. It contains the whole source code of the platform and add-on features, and several *readme* files describing the different import and test functions to be run. The detailed documentation and Github are not made publicly available. The DSpace platform itself is open source, but the customisation made in this project to create the platform are specific to our use case and to the data we integrated in it. In addition, security information necessary to operate the platform and the database are present in the code. In addition, the whole server containing the platform is part of a nightly backup routine within the NUI Galway infrastructure. At the time of this submission, there is not yet a public documentation on how to use the platform, but this will be included in the following weeks at the official release.

We have described in this section all the features available in the Linghub2 platform for search and discovery of data. We will now present additional processes developed to ensure quality and standardisation of the data.

5. Quality Tests and Conversions

Some of the main goals of this platform are to provide harmonized metadata that is easy to find, easy to access and use, and allows the user to know the quality of the data they are looking at. Tests are performed to validate the quality of the data and of the LLOD standards agreed by the Linguistic Linked Open Data community whose members are part of the consortium of this project.

5.1. Use of Standards

Metadata schemas of resources typically come in heterogeneous formats, depending on the source they are collected from. Homogenisation of schemas was already a priority in the first version of Linghub, as harmonisation of formats is needed for an efficient search of the data for standard metadata such as title, language, subject. Conversion scripts to the standard DC and DCAT¹³ schemas were created for each source (repositories) in the first version of Linghub, and reused in Linghub2 to convert the updated data before import. Where metadata schemas and fields could not be converted as there was no equivalent in DC or DCAT, the metadata was still added to the platform as additional information to keep as much information as possible. Specific scripts were developed for the conversion from the RDF files that we collected into the CSV format required by DSpace, containing the metadata as Linked Data (LD).

⁸<https://solr.apache.org/>

⁹<https://www.iso.org/iso-639-language-codes.html>

¹⁰<https://www.w3.org/TR/rdf-sparql-query/>

¹¹<https://jena.apache.org/documentation/fuseki2/>

¹²<https://github.com/>

¹³<https://www.w3.org/TR/vocab-dcat-2/>

Search

The search interface shows a search bar with 'speech conversations' and a 'Go' button. Below the search bar, there are two filter tags: 'Language (ISO): zul' and 'Type: sound'. A 'Hide Advanced Filters' link is visible. The 'Filters' section includes dropdowns for 'Language ()' and 'Type', and input fields for 'zul' and 'sound'. There are also 'Reset' and 'Apply' buttons. Below the filters, it says 'Now showing items 1-1 of 1'. A 'No Thumbnail' placeholder is shown next to the search result for 'IARPA Babel Zulu Language Pack IARPA-babel206b-v0.1e'. The result description lists authors: Lin, Willa; Dubinski, Eyal; Wong, Jamie; Fiscus, Jonathan G.; Adams, Nikki; Connors, Thomas; Silber, Ronnie; Harper, Mary; Ray, Jessica; Tzoukermann, Evelyne; Bills, Aric; Melot, Jennifer; Rytting, Anton; Shen, Wade.

Figure 3: Searching for Zulu Spoken Conversation using the Advanced Search

The 'DISCOVER' section shows a list of languages and policies. The 'Language (ISO)' section includes: nld (20639), fra (13069), eng (12871), deu (11703), jpn (4460), and tur (3817). The 'Policy' section includes: http://purl.org/NET/rdflicense/cc-by-nc-sa3.0 (6469), http://purl.org/NET/rdflicense/cc-by-nc-nd3.0 (4165), http://purl.org/NET/rdflicense/cc-by-nc-nd4.0 (660), http://purl.org/NET/rdflicense/PDM1.0 (483), and http://purl.org/NET/rdflicense/publicdomain (480).

Figure 4: Discovery Feature

The SPARQL endpoint interface shows a query:

```
1 PREFIX dcl: <http://purl.org/dc/elements/1.1/>
2 PREFIX dcterms: <http://purl.org/dc/terms/>
3
4
5
6 SELECT ?subject ?predicate ?object
7 WHERE {
8   ?subject ?predicate ?object
9 }
10 LIMIT 25
11
```

 The results table shows three rows of data with columns for subject, predicate, and object.

Figure 5: SPARQL endpoint and query

In order to allow the user to easily identify data that follows standards established by the Prêt-à-LLOD community, an icon is displayed with a label in the platform when the metadata from the resource is using standard metadata vocabularies only (tick) or not (cross).

5.2. URL Verification

We developed and integrated a URL health check module that verifies the availability of URLs given

in the resource description. These URLs link to either the source repository resource description, the page where the resource can be downloaded, the main website URL of the organization providing the resource, etc. For this, a list of metadata fields related to where data can be accessed or downloaded (ie. *dcat.accessURL*, *dcat.endpointURL*, *dcat.downloadURL*, *foaf.homepage*, *dcat.landingPage*) is checked to verify their availability using HTTP sta-

tus code. This health check is implemented as a periodic curation task executed using cron jobs to keep the health check flags up-to-date. A ‘tick’ or ‘cross’ mark is added in the UI with a label to allow the user to easily know at a glance whether the data they are looking at is available or not (see Figure 6 below).


rdftype	http://www.w3.org/ns/dcat#Distribution
dcat.accessURL	http://zhishi.me/sparql  URL may not be accessible
dcat.mediaType	api/sparql

Figure 6: URL Test Feature

5.3. Language Mapping and Conversion

Language values can be quite inconsistent across resources, which makes it difficult for the user to know how to search for a particular language. We developed a language mapping module that maps language-related metadata fields from the resource to standard ISO 639 values, and adds a specific field in the metadata (*dc.language.uri*) when a language is recognised, ie. lexvo URLs (de Melo, 2015), ISO 639 code, or the corresponding language name of the ISO code (eg. *English*). It is part of the dataset import module and is executed in the pre-processing pipeline. A ‘tick’ or ‘cross’ mark is also added in the UI with a label to flag the language values that are extracted in this manner (see Figure 7).

5.4. ODRL Conversion and Query

Understanding of the scope of a dataset license is a step towards ensuring the correct usage of LRs and technologies by users. The Open Digital Rights Language (ODRL)¹⁴ is an official W3C recommendation, and a known standard for modeling all types of licenses and policies. This complex policy expression language is developed and provided by the Polytechnic University of Madrid. By working closely with the ODRL community and focusing on this standard, our goal was to ensure license quality in the metadata, and help non-experts to understand whether or not they can use the dataset for their purpose. In order to do so, we produced a number of developments around the ODRL.

5.5. ODRL API development

An API to facilitate the operations related to licenses was designed, developed and deployed as part of PAL. The API operates over a dataset of commonly used licenses (such as the Creative Commons, or Apache licenses), and an arbitrary set of policies represented in ODRL using the ODRL for LRs profile. The ODRL representations of these licenses was generated and published to an open git repository, using an initial version of the data model under development within the

¹⁴<https://www.w3.org/TR/odrl-model/>

W3C ODRL Community Group. An HTTP REST API was also developed and documented using Open API standards (Swagger), and is available online¹⁵.

5.6. ODRL conversion module

For data quality purpose, we converted license-related metadata that we could identify from the resources (ie. expressed using standard denomination such as “Attribution-ShareAlike 2.0”, or “CC BY-SA 2.0”), to their corresponding ODRL identifier and added them in separate metadata field: *odrl.Policy*. A limited number of licenses are currently described with the ODRL language (all the main creative commons and a few others), but this work is in constant evolution. A ‘tick’ or ‘cross’ mark is also added in the UI with a label to flag the ODRL policies that are automatically extracted in this manner (see Figure 8).

5.7. ODRL form

Ensuring the lawful use of data and language technologies is a challenge. Although we can not enforce the right usage of the data, we can still help with providing non-specialists a non technical explanation of the implications of a license. For this purpose, we provide a form within the platform that allows the user to verify whether or not they can use the data for their purpose, based on the ODRL information present in the metadata and using the ODRL API.

The form first asks about the purpose of the use of the dataset, the affiliation of the user and the duration for which the dataset is needed (the user can select a date, or leave blank for unlimited). The form has a limited number of options at the time of submission, but research collaborations within UPM are aiming at extending them in the future. Note that such a form validation feature is only given for the datasets for which a mapped ODRL policy was identified in the metadata. The form is shown on the main display page of the resource (see Figure 9 for an example)

We have now described the main features of the platform, and we will explore in the next section the data sources used to populate the Linghub2 platform, and what has changed since the first Linghub version.

6. Sources and Statistics

We describe here the resources present in the Linghub2 platform. We have set up a minimum requirement for the resource to have at least metadata on the title, to filter out datasets that are too obscure. DSpace organises datasets in terms of *Collections*, which correspond to data sources. The following collections are currently present: CLARIN, OLAC, old.datahub, LRE Map, META-SHARE, Annohub, Teanga, iLOD. Some repositories were present in the first Linghub and were updated, and some new data sources were added in this second version.

¹⁵<https://rdflicense.linkeddata.es/>

dc.description	Swadesh wordlist revised by Blust and some additional lexical items. Language as given: Lio
dc.format	Digitised: no Media: digital
dc.language	http://lexvo.org/id/iso639-3/lji
dc.language.iso	lji
dc.publisher	http:// This field was automatically generated based on the metadata provided. It follows Prêt-à-LOD standards recommendations

Figure 7: Language Test and Conversion Feature

odrl.Policy	http://purl.org/NET/rdflicense/cc-by-nc-sa3.0
dc.bibliographicCitation	http://hdl.handle.net/11858/00-097C-0000-0005 This field was automatically generated based on the metadata provided. It follows Prêt-à-LOD standards recommendations

Figure 8: ODRL Test and Conversion Feature

Check resource access

Choose a purpose

Who are you?

When

Authorized	No
Reason	This resource cannot be used for the given purpose.

Figure 9: ODRL Rights Checking Form

6.1. Existing Data Updated

The lack of interoperability across datasets is partly linked to the issue of developing specific schemes. Many domain specialist groups and organizations have developed their own representation of data, that fits their research and objectives. As valuable as this motivation is for the particular group, this can be an impediment for the larger community and non domain experts, in the access, use, and discovery of data. This is the case of CLARIN (Van Uytvanck et al., 2012) (with the CMDI schema (Broeder et al., 2012)) and META-SHARE (Gavrilidou et al., 2012) which have both defined their own scheme. McCrae et al. (2015) dedicated significant work into mapping schemes and building specific conversion tools for these repositories in Linghub (first version), in order to tackle the problem of heterogeneity of schemes, and also format (the repositories are not using LD, but XML formats instead). Since then, we collected the latest version of the data available, and reused these tools for Linghub2. The META-SHARE resources in Linghub2 were provided to us by the European LR Association (ELRA). The anonymised META-SHARE XML format dump of the data was extracted on 17/06/2021. For CLARIN, since the CMDI schema was updated since, a new har-

vester was recently developed. 2,764,055 resources were extracted, and are being uploaded into the platform at the time of writing. Changing schemes are a limitation for the development of LRs and for users who need to keep up-to-date with the scheme.

old.datahub.io¹⁶ on the other side, is a CKAN-based platform¹⁷, which means it uses standard LD schemas and can be easily crawled, using the same harvester as in Linghub version 1. The LRE Map¹⁸ platform (Calzolari et al., 2012) website however is not based on any repository management system and does not follow standard Web design for platforms. It has proven impossible to update the harvester from Linghub version 1 to tackle the source code changes of the Website, restraining us to use the old 2014 data.

6.2. New Data

OLAC. The Open Language Archives Community (OLAC) (Bird and Simons, 2003) provides a virtual library of LRs¹⁹. They have developed their own LD-based metadata set²⁰, however it is based on the complete set of the standard DC schema. Since they are using LD, we were able to harvest this new collection in Linghub2 through OAI-PMH.

Annohub. The Annohub dataset (Abromeit et al., 2020) was created and provided by our project partner Goethe University Frankfurt. It is an open license dataset, containing annotated LRs like corpora freely available on the internet, enriched with automatically generated metadata using an automatic workflow followed by a curation by domain experts. In the future, the workflow will serve Linghub2 to augment its resources with missing information, for example lan-

¹⁶<https://old.datahub.io/>

¹⁷<https://ckan.org/>

¹⁸<https://lremap.elra.info/>

¹⁹<http://www.language-archives.org/>

²⁰<http://www.language-archives.org/OLAC/metadata.html>

guages.

Teanga services. Teanga²¹ (Ziad et al., 2018) is a LD-based platform for NLP that enables the use of many NLP services from a single interface, as a workflow. Services compatible with Teanga (either initially or converted) were provided through the PAL project. The initial set comprises 47 services, providing ready-to-use software components for NLP.

iLOD. The iLOD dataset (Nasir and McCrae, 2020) was created as part of the Prêt-à-LLOD project, containing data from the LOD cloud available through the InterPlanetary File System (IPFS).

These datasets will be easy to update in the future as they are available in LD standards.

6.3. Statistics

In Table 1 below is a summary of the data currently available in Linghub2, and the number of languages covered by each repository (as detected by our language mapping module). There has been significant changes in the number of available data, as can be seen from the original data of Linghub, in Table 2. We note that these numbers will include duplicates in the data, as we have not been able to run our duplicate detection module (see section 7). As part of the KPIs of the Prêt-à-LLOD project, the platform aimed at reaching 1,000,000 datasets on offer (as opposed to the 150,000 in the first version of Linghub), the 24 EU languages, and 300 other languages.

Source	# Language Resources	# Languages
OLAC	442,501	645
META-SHARE	107,474	160
CLARIN	2,764,055	49
old.datahub	2,615	0
LRE Map	1,455	0
Annohub	615	2,760
iLOD	90,598	0
Teanga	47	0
Total	3,409,360	3,614

Table 1: Resources and languages in Linghub2

Source	# Language Resources
META-SHARE	2,442
CLARIN	144,138
Datahub	185
LRE Map	682
Total	147,447

Table 2: Resources statistics in Linghub version 1 (McCrae and Cimiano, 2015)

²¹<https://teanga.io/>

The platform has reached a number of achievements, however some functionalities had to be dropped. We detail now some challenges faced during the development of the platform.

7. Discussion and Challenges

The new Linghub2 platform is providing a more modern looking interface, with many features built in the data management software, allowing robustness and ensuring sustainability.

Most of the key challenges for the new platform were reached, however some were out of reach in terms of time, or feasibility.

Third party software use. The version of DSpace installed for Linghub is currently DSpace 6.x, which was the latest stable version available at the time of development. However, DSpace 7 was released in August 2021. In this newest version, which is the largest release in the history of the DSpace software, sees major changes with many features re-imagined and re-implemented, including a new UI. DSpace 6 will therefore not be further developed, and a transition to DSpace 7 will be necessary in the future, which could lead to substantial work.

Embedded metadata. Some entries in Linghub2 contain subsets of resources. In these cases, some of the metadata apply to the entire resource, and some are more specific to each subset of the resource. Unfortunately, DSpace 6 does not support hierarchical metadata (this is tackled in DSpace 7 however). The platform therefore only displays a flat structure, but not the hierarchy of the metadata of the subsets. Not having full control over the functionalities is the inconvenience of working with a third party system, however this was a compromise that we accepted. Since the metadata displayed also contains the original data source URL, this missing feature does not impact the data discovery capability which is the main aim of our tool.

Data update. The original plans included automating the update of the data from the repositories present in the platform. However during the course of the project we identified several blockages. Most repositories have developed their own data schemes, and the conversion modules for Linghub2 are built specifically for them. Their schemes also typically evolve with time, as the communities work further on theme, which makes it necessary to update conversion modules accordingly. Finally, each data collection was collected through different channels, from the community responsible for it (META-SHARE), from their website (OLAC), from crawling the repository online, etc. For each of these cases, there is a likelihood that the source has changed and/or is not available anymore (eg. LRE Map website structure has changed

and is not standard, making it impossible to crawl the repository). This implies that it is impossible to rely on an automated process for updating resources. Moreover we do not allow users to add their data to the platform by themselves, as there is a number of quality criteria (such as the standardisation of the schemas) and tests that we want to control. However, we will provide a documentation on different ways to provide the data (ie. as a file that we convert into the right DSpace format, or in the DSpace format directly).

Data Quality. In the initial plans as well, Linghub2 was providing a verification of not only the metadata, but the resource data itself. Identifying badly formatted XML files for examples, or data not accessible for download. However, this implied downloading each dataset available from each resource in Linghub2. Some resources are readily accessible through a direct link, others redirect users to the main website page, while others require users to create an account, and others require a payment. Between these numerous cases and the huge processing memory needed to download and parse datasets, we instead created the URL check module (see 6), that tells whether a link is valid or not.

Duplicate Detection. We have designed and developed a duplicate detection module, that allows only one resource to be displayed in the search results if other duplicates exist (the one containing the most metadata). It also gives links to the other duplicates when displaying the individual resource. The module compares the values of the *dc.identifier* and *dc.identifier.uri* metadata for all resources of the platform. However, the runtime over the whole data has proven to be very high, and therefore not realistic to run.

8. Conclusion

In an attempt to reduce the amount of time spent by language technology specialists in searching and preparing relevant language data and to reduce the loss of LRs that are not centralized and standardized, we developed Linghub2, a platform built on a pre-existing data discovery tool, Linghub. The new Linghub2 platform has increased the data content of the previous Linghub significantly, as well as the languages covered. It features a user-friendly and easy to use interface for less technical users, as well as powerful search capabilities, and is based on a robust open source software which is widely used.

While the current platform has successfully improved in terms of user experience, robustness, data content, quality and standards, some challenges remain. The heterogeneous formats and accessibility of the data at the source is still an impediment for automated solutions of integration and standardisation. Moreover, relying on a third party software can also bring limitations, and the maintenance and development of the tool

in the future is directly dependant on the support it receives, which we cannot control.

9. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182, and SFI/12/RC/2289.P2 (Insight), co-funded by the European Regional Development Fund.

10. Bibliographical References

- Bird, S. and Simons, G. (2003). Extending dublin core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 37:375–388, 01.
- Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.
- Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE Map. harmonising community descriptions of resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, page 1084–1089. European Language Resources Association (ELRA), 05.
- de Melo, G. (2015). Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*, 6:393–400, 05.
- Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., Mapelli, V., and Ilsp, A. (2012). The META-SHARE metadata schema for the description of language resources. page 1090–1097, 05.
- Khan, S. (2019). Dspace or Fedora: Which is a better solution? *SRELS Journal of Information Management*, 56(1):45–50, 02.
- McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In *SEMANTiCS*.
- McCrae, J. P., Cimiano, P., Rodríguez-Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015). Reconciling heterogeneous descriptions of language resources. In *LDL@IJCNLP*.
- Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajič, J., Hamrlová, J., Kačena, L., Choukri, K., Arranz, V., Vasiljevs, A., Anvari, O., Lagzdinš, A., Meļņika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Gómez-Pérez, J. M., Garcia Silva, A., Berrío, C., Germann, U., Renals, S., and Klejch, O. (2020). European

language grid: An overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France, May. European Language Resources Association.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the virtual language observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ziad, H., McCrae, J. P., and Buitelaar, P. (2018). Teanga: A linked data based platform for natural language processing. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

11. Language Resources References

Abromeit, F., Fäth, C., and Glaser, L. (2020). Annohub – Annotation Metadata for Linked Data Applications. *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 36–44, May.

Nasir, J. A. and McCrae, J. P. (2020). iLOD: InterPlanetary File System based Linked Open Data Cloud. *Proceedings of MEPDaW20 - Managing the Evolution and Preservation of the Data Web, ISWC 2020*, Nov.