



D5.3 Prêt-à-LLOD Language Resource Discovery Portal

Author(s): Cécile Robin, Gautham Suresh,
Jamal Nasir, Bernardo Steams, John
McCrae, Paul Buitelaar, Víctor Rodríguez

Date: 10th September 2021



H2020-ICT-29b

Grant Agreement No. 825182

Prêt-à-LLOD - Ready-to-use
Multilingual

Linked Language Data for
Knowledge

Services across Sectors

*D5.3 Prêt-à-LLOD Language Resource
Discovery Portal*

Deliverable Number: D5.3

Type: Software

Dissemination Level: PU

Delivery Date: 10/09/2021

Version: 1.2

Authors: Cécile Robin, Gautham Suresh, Jamal Nasir, Bernardo Steams,
John McCrae, Paul Buitelaar, Víctor Rodríguez

Document History

Version	Date	Changes	Authors
1.1	Sep 2, 2021	Internal Review	Matthias Hartung(SEM)
1.2	Sept 10, 2021	Corrections after review	Cécile Robin (NUIG)

Table of Contents

Overview	4
Meeting the objectives	5
Challenge 1	5
Work plan	7
Linghub - DSpace-based platform	8
DSpace Software	8
Technologies	9
Features	10
Linghub customization	12
Data Quality Preprocessing	12
URL health check	13
Language mapping	13
ODRL Policy mapping	13
UI customizations	13
Home page	13
Resources display	14
URL verification	15
Language mapping	15
ODRL Policy mapping	15
Teanga service identification	
An icon is added to a resource (near its title) in the UI to mark if the resource is a service compatible with Teanga.	17
Advanced search	17
ODRL API and query form	18
API	18
Query form	21
SPARQL endpoint	22
Resources	23
CLARIN	23
Overview	23
CLARIN in Linghub	24
OLAC	24
Overview	24
OLAC in Linghub	25
old.datahub	25
Overview	25
old.datahub in Linghub	25
iLOD	25



LRE Map	26
Overview	26
LRE Map in Linghub	26
META-SHARE	26
Overview	26
META-SHARE in Linghub	26
Annohub	26
Teanga services	27
Linked Data Sustainability through IPFS	27
Maintenance	28
Conclusions	28
Technical Documentation	28

Overview

This document reports on the Linghub language resource discovery portal, the software deliverable developed as part of task T5.3 (“Repositories for Resources and Metadata”) which concludes the delivery of work package 5 in M37 (July 2021), with one month delay to the original plan. The task fits in the challenge 1 described in the project proposal, which is about the discovery of language resources, using homogeneous and expressive metadata from different language resource repositories, giving information about availability, technical quality and content of language resources and combining them into a single search interface.

The objectives linked to the challenge and work plan were met through the delivery of the platform. We will describe first these objectives and main points of the work plan, and how they are tackled in the platform. We then present the software which Linghub is based on, and the customizations that were made to bring the platform up to the standards desired. Next, we list the resources harvested and imported in Linghub, coming from different repositories and sources. Then we describe the exploration of an approach to the sustainability of data through a peer-to-peer decentralized storage infrastructure and the iLOD dataset created as a result of this task. Finally, we touch on the long term remaining tasks concerning the maintenance of the platform.

1. Meeting the objectives

1.1. Challenge 1

This task is part of challenge 1, described in the project proposal: “**Discovery** of language resources across multiple repositories. Methods for linking metadata repositories will be developed and combined into a single search interface.”

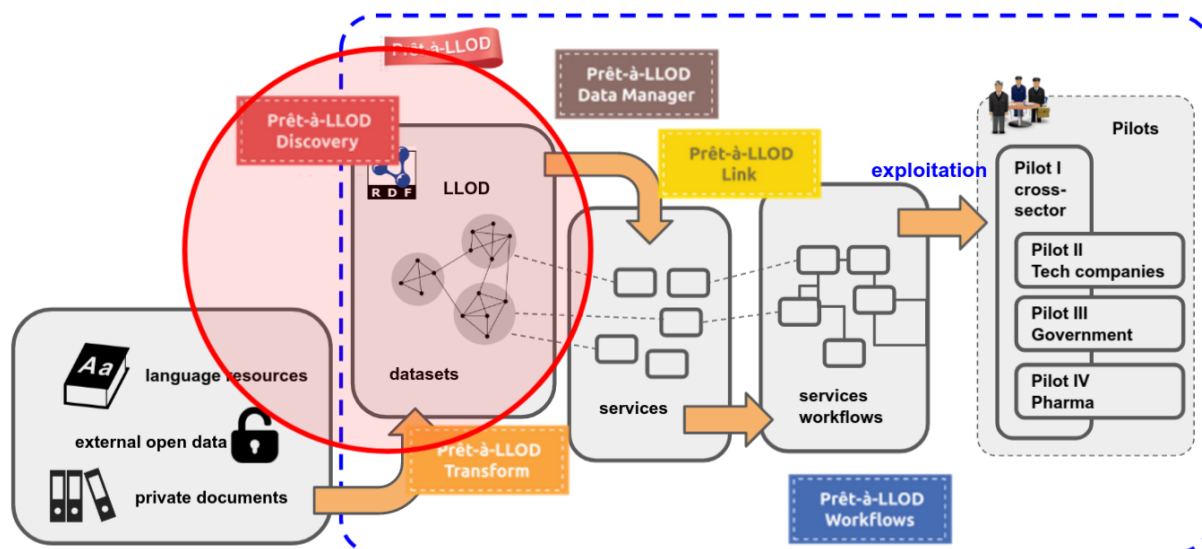


Figure 1: Prêt-à-LLOD pipeline

Multiple areas of focus were part of this challenge, such as sustainability through accessibility and repository management, long-term storage, and providing access to the data and related services, all this by building on the previously developed Linghub portal and engaging with the language resources communities (in particular the Linguistic Linked Open Data community). These key requirements were tackled all throughout the duration of the task.

We refactored the previous Linghub website by using a widely used open source data software platform, DSpace¹, that utilises well-established technologies for data management, storage and search (tomcat, solr, postgresSQL). By relying on software developed specifically for this purpose and continuously maintained and improved over time by an ever-growing community of developers, we guarantee the sustainability of the platform. This solution ensures open access and long term management (preservation) of the repository. The source code of the platform is stored in the Prêt-à-LLOD (PAL) Github's repository,² which ensures openness, community work and version control.

This new version of Linghub provides a multilingual cross-repository data search facility, and covers major dataset sources across Europe and the world including DataHub³, and in particular language resource repositories including LRE Map⁴, META-SHARE⁵, CLARIN⁶ and more (more detail is given in section 4). This allows for open access to heterogeneous metadata coming from distinct metadata repositories, all searchable in one portal with standardized metadata.

¹ <https://www.lyrasis.org/DCSP/Pages/DSpace.aspx>

² <https://github.com/Pret-a-LLOD>

³ <https://old.datahub.io/>

⁴ <https://lremap.elra.info/>

⁵ <http://www.meta-share.org/>

⁶ <https://www.clarin.eu/>

An extensive engagement with the language resources communities was made throughout the development of the platform, and is reflected at different levels. In the platform we introduced tests and conversions of linked data schemas used in the harvested metadata, linking back to the standards established by the community in the outcomes of T5.1, detailed in D5.1 - Report on Vocabularies for Interoperable Language Resources and Services.

A data model for licenses used in the language resources domain was specified as a profile of the Open Digital Rights Language (ODRL), which is a W3C Recommendation to represent computer policies including permissions, prohibitions and obligations. A collection of licenses commonly used in this domain was provided and is being served. Services to perform reasoning tasks with licensing information were developed, including operations such as access, validation, authorization and license compatibility checks .

The platform links to the consortium's effort in terms of standardization tools, as the services developed by the partners throughout the project are and will be continuously added in the platform. The workflow of Teanga is also connected, as Teanga compatible services are specifically marked in the platform.

Finally, sustainability in the data is ensured through the development of the iLOD dataset (described in more details section 4.4 below), a dataset created using Blockchain-based InterPlanetary File System (IPFS) to preserve the Linked Open Data Cloud.

1.2. Work plan

In relation to the work plan, we are providing a new version of Linghub with a harmonized access to language resource metadata with standardized metadata control. Three key aspects were part of the work plan:

i) **availability**: In this work, we perform a url check of metadata related to web pages and resource access (dcat.accessURL, dcat.endpointURL, dcat.downloadURL, foaf.homepage, dcats.landingPage). By using a visual flag on the resource display, we let the user know whether the links are broken or not, helping them identify which resources are accessible or not.

ii) **quality**: The quality of the data provided is addressed by the use of standard metadata schemas. For example, data coming from META-SHARE and CLARIN were converted to Dublin Core and dcat schemas, while the Open Language Archives Community (OLAC⁷) data was harvested through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which by default retrieves Dublin Core metadata. Moreover, we display a visual flag that alerts the user when other schemas are used that are not recommended by the community in T5.1 (Dublin Core, DCAT, ODRL, etc). This allows the user to know at a first glance whether the metadata they are looking at is following quality standards or not. Moreover, through the work on the iLOD dataset using IPFS, we are pushing the standardization and sustainability of data (see more on this in section 5)

⁷ <http://www.language-archives.org/>



iii) **content**: On the content side, most of the data previously present on the Linghub was updated to the latest version available from the repositories, and more data was added through a new generated corpus, iLOD, based on data from the LOD cloud. We also extended the available metadata by automatically generating the ISO version of the language of the resource, as well as the ODRL policy it was mapped to (on resources for which we could identify this information). This ensures both quality and standardization of this very important information available in the content.

In this task we have provided a single harmonized interface for enabling users to find metadata from various resource sharing and metadata sources. The sources contain both language resources and services, which will be updated through time. By using well established technology for data management, we are ensuring the sustainability of the platform and the data it contains. The platform also contains a data search facility using Solr⁸. Where available, we are using the ODRL schema to represent policies. We also developed a form through which a user can verify their right to use the resource or not. This way, we help ensure the lawful use of data and language technologies.

The results of this analysis will be made available and searchable through the platform at <https://linghub.org> (currently available at <http://new.linghub.org/>), and a SPARQL and faceted search is also offered.

Finally, the source code of the platform is saved in a git repository, ensuring sustainability and version control: <https://github.com/Pret-a-LLOD/Dspace/>. This repository is not public as it contains sensitive information about the platform, ie. configurations with password.

2. Linghub - DSpace-based platform

2.1. DSpace Software

For this project, a refactoring of the previous Linghub platform was motivated by the provision of a platform for storing data that is sustainable and can be easily deployed by specialists in the Linguistic Linked Open Data domain without the need for a deep knowledge of web development, while having characteristic features of such a tool made easy to implement.

Therefore, it was agreed to change the previous version of Linghub to a data management software developed specifically for this type of use. Several open source tools exist. CKAN⁹ was first considered and recommended by some partners, however the difficulties to set it up and complexity of the use of the platform for non-experts stopped us from going in this direction. Instead, DSpace¹⁰ was chosen, another widely

⁸ <https://solr.apache.org/>

⁹ <https://ckan.org/>

¹⁰ <https://www.lyrasis.org/DCSP/Pages/Dspace.aspx>



used and open-source platform developed and maintained by Lyrisis and their extended community.

DSpace is a software for building open digital repositories used by academic, non-profit and commercial organizations. It is open source (Creative Commons Attribution 4.0 International License), and developed and supported by a strong user community, with the help and guidance of DuraSpace/Lyrisis. It is widely used, with more than a 1000 (known) instances of DSpace running worldwide, and is constantly evolving. It is supported by different organizations, such as ConcyTec, Cambridge University Library, Cornell University Library, Imperial College London, etc.

The version of DSpace installed for Linghub is currently DSpace 6.x, the latest stable version available.

DSpace is aimed at being free, easy to install (“out of the box”), and customizable to adapt to different organizations’ needs. DSpace is designed to allow easy and open access to all types of digital content including text, images, moving images, mpegs and data sets, although in our case with Linghub we are only interested in the metadata of such resources.

DSpace has an ever-growing community of developers that continuously work on expanding and improving the software and many options are available to get support. A detailed documentation is provided¹¹ and maintained with the newer additions and changes, and a user FAQ¹² and technical FAQ¹³ are also available. The community can also be contacted via different means (mailing list, Slack channel). Moreover, a number of answers can also be found on Stackoverflow¹⁴ with the tag “DSpace”. The full list of all resources of support is given on their web page¹⁵.

2.2. Technologies

DSpace is using different well established and open source tools to operate.

It uses Apache Tomcat®¹⁶ to power the web application. It is an open source implementation of the Jakarta Servlet, Jakarta Server Pages, Jakarta Expression Language, Jakarta WebSocket, Jakarta Annotations and Jakarta Authentication specifications. The Apache Tomcat software is developed in an open and participatory environment and released under the Apache License version 2.

DSpace relies on Apache Solr a popular and scalable open source enterprise search platform built on Apache Lucene™. It provides distributed indexing, replication and load-balanced querying, and powers the search and navigation features of the web application. Solr is a reliable technology used by some of the world's largest internet sites.

¹¹ <https://wiki.lyrisis.org/display/DSDOC6x/DSpace+6.x+Documentation>

¹² <https://wiki.lyrisis.org/display/DSPACE/User+FAQ>

¹³ <https://wiki.lyrisis.org/display/DSPACE/TechnicalFAQ>

¹⁴ <https://stackoverflow.com/questions/tagged/dspace>

¹⁵ <https://wiki.lyrisis.org/display/DSPACE/Support>

¹⁶ <http://tomcat.apache.org/>



On the data storage side, DSpace is using postgresSQL¹⁷, an open source relational database system with a strong reputation for reliability, feature robustness, and performance.

By integrating all these robust technologies into its design, DSpace ensures a stable and sustainable data management system.

More information on the installation of DSpace for Linghub and the versions of these tools installed is provided in the Technical Documentation in the annexe.

2.3. Features

DSpace provides in its design a wide number of features desired in such a platform.

Search / Browsing / Filtering: Thanks to Solr, DSpace provides a built-in search facility that allows the user to easily search for specific content within the data available. An advanced search allows the user to specify in which collection they want the search to be performed, as well as specifying the metadata field on which they perform the search (figure 2).

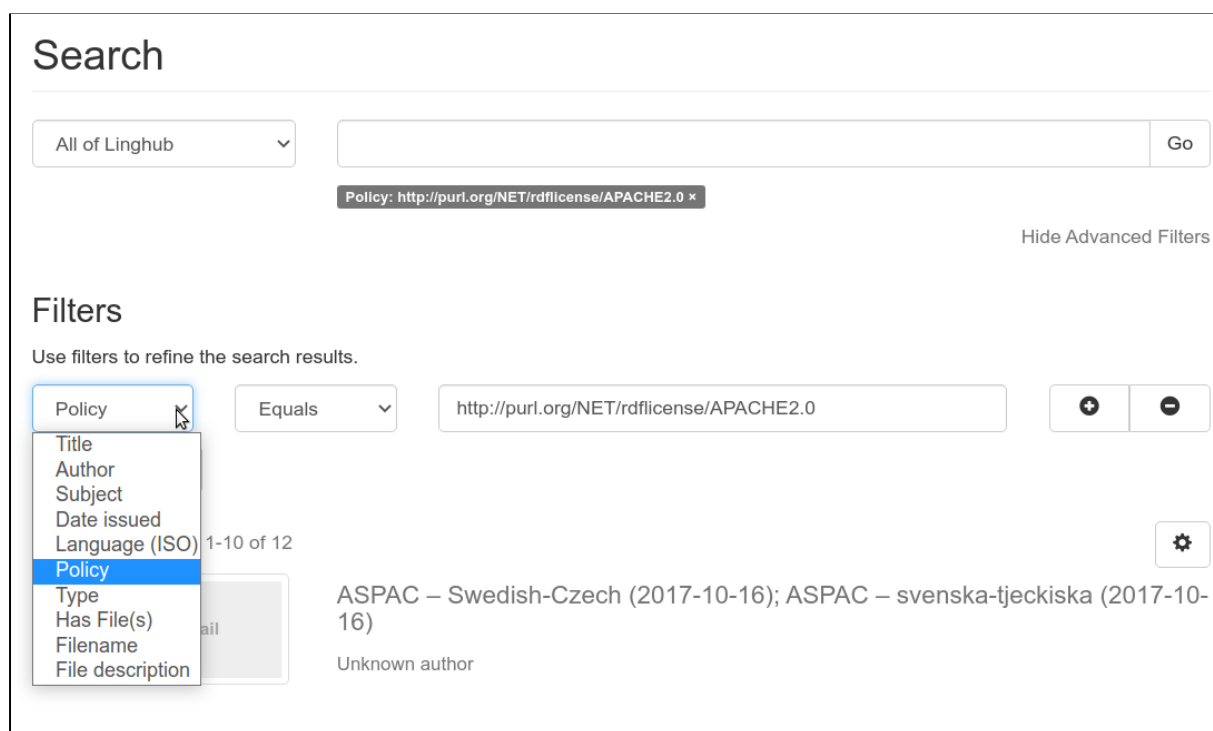


Figure 2: Advanced search

It is also possible for the user to simply browse and discover data according to a specific metadata field, using a sidebar displaying a selection of metadata available in the resources. See an example in figure 3 below.

¹⁷ <https://www.postgresql.org/>

DISCOVER	
Author	
Christian Chiarcos	(379)
Kevin Scannell	(76)
Gilles Sérasset	(50)
Laboratoire Ligérien de Linguistique	(47)
Biggs, Patricia	(24)
Blanc, Michel	(24)
Faculté de Linguistique Appliquée de l'Université d'Etat d'Haïti, (anciennement Centre de Linguistique Appliquée (CLA))	(17)
Maddieson, Ian	(17)
Henry Kučera	(16)
W. Nelson Francis	(16)
... View More	
Subject	
enquête	(4631)
Français	(1859)
http://lexvo.org/id/iso639-3/und	(1468)
Leichenpredigt	(1366)
Dissertation:jur.	(944)

Figure 3: Sidebar for discovery of resources

Import: DSpace allows different types of batch import of data into the platform, one of them by using their own format called Simple Archive Format (SAF)¹⁸, or other bibliographic formats (Endnote, BibTex, RIS, TSV, CSV) and online services (OAI, arXiv, PubMed, CrossRef, CiNii)¹⁹. We found that using CSV was the most simple, flexible and complete way to import large quantities of data coming from different sources. This was therefore our chosen way to import metadata in batches.

REST API: DSpace also comes with a REST API, where it is possible to query the metadata schemas, fields and value of complete collections or individual items. It also allows modification and import of existing and new data. The REST API is only available for authenticated users registered on the Linghub platform; therefore we will be restricting this usage for a selected number of providers. We are using this API to collect information on the metadata (such as id, metadata schemas) for the various metadata validity tests, or to import data into the platform, for direct maintenance of datasets in collections where the data was created from the Pret-a-LLOD project and is in constant development (ie. the collections related to the iLOD dataset and the Teanga compatible services, explained in section 4)

¹⁸

<https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format>

¹⁹ <https://wiki.lyrasis.org/pages/viewpage.action?pageId=45548176>

Linked Data-based metadata: The metadata stored in DSpace is using Linked Data, more specifically by default the Dublin Core schema, although it is possible to add different schemas and fields.

Cron tasks²⁰/curations tasks²¹: It is possible to set up (cron) tasks to be run regularly following a specific schedule. Some curation tasks are available “out of the box” in the platform, and it is also possible to create personalized tasks. **These tasks are helpful, for example to regularly test URLs in the metadata, adapted from the pre-set “check-links” task available in DSpace.**

Privacy/authentication: DSpace allows users to be added to the platform, with defined roles such as contributor, administrator, etc., each role having a specific set of permissions. It is also possible to restrict the access of certain data to a specific group of users.

Backup: DSpace includes a capability to backup the platform and its data to either a local storage, a mountable storage, or a DuraCloud Storage, using AIP package format.²²

Linked Open Data²³ / SPARQL endpoint: DSpace supports publishing stored contents in the form of Linked (Open) Data. It enables data conversion from the platform into RDF and storage in a triple store immediately after creation or update. The triple store serves as a cache and provides a SPARQL endpoint to make the converted data accessible using SPARQL.

3. Linghub customization

Some customizations were applied to the standard DSpace platform to create the final version of Linghub that matches our criteria in terms of quality, sustainability, esthetics, and ease of use, both in the back-end as well as in the front-end.

The platform is available at the address: <https://linghub.org> and the source code in <https://github.com/Pret-a-LLOD/DSpace/> (private access since it contains confidential information).

3.1. Data Quality Preprocessing

One major focus for this platform is to provide metadata that is harmonized and easy to find. Validation tests are performed, as well as mappings to Linked Data metadata standards agreed by the Linguistic Linked Open Data community, as outlined in D5.1 (Dublin Core standard as opposed to META-SHARE specific schema for example).

²⁰ <https://wiki.lyrasis.org/display/DSDOC6x/Scheduled+Tasks+via+Cron>

²¹ <https://wiki.lyrasis.org/display/DSDOC6x/Curation+System>

²² <https://wiki.lyrasis.org/display/DSDOC6x/AIP+Backup+and+Restore>

²³ <https://wiki.lyrasis.org/display/DSDOC6x/Linked+%28Open%29+Data>



3.1.1. URL health check

The URL health check module verifies the availability of URLs given in resources. For this, a list of metadata fields will be checked to verify their availability (using HTTP status 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 show the health check values to allow the user to easily know at a glance whether the data they are looking at is available or not.

3.1.2. Language mapping

A language mapping module was developed to map language-related metadata fields from the resource to standard ISO values, and to add a specific field (named *dc.language.uri*) to the existing resource metadata when a language is recognised. 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.

3.1.3. ODRL Policy mapping

A policy mapping module was developed to map values from licensing-related metadata fields in the resource to ODRL policies, and to add a specific field (named *odrl.Policy*) to the existing resource metadata when a policy is recognised. It checks the field values for license names, license codes, PURL and actual URL for the licenses, based on a set of standard licenses converted to ODRL, made available by project partner UPM²⁴. A ‘tick’ or ‘cross’ mark is also added in the UI with a label to flag the ODRL policies that are extracted in this manner.

3.2. UI customizations

A series of customizations were applied on the front end of the platform to match the Linghub/Pret-a-LLOD theme and display the resources made available.

3.2.1. Home page

The color scheme of Prêt-à-LLOD with the logo, as well as all the consortium partners logos and H2020 funding were added to customize the default DSpace UI, as can be seen in the figure 4 below.

²⁴ <https://github.com/Pret-a-LLOD/pddm/tree/develop/data/licenses>

Prêt-à-LLOD Login

Linghub Home

Linghub

A comprehensive location for finding information about language resources. We take records from different sources and make them available using RDF linked data, DCAT and SPARQL. We use state-of-the-art technology to process different metadata records and make all the data available under a common scheme.

Communities in Linghub

Select a community to browse its collections.

- Annohub
<https://annohub.linguistik.de/>
- CLARIN
<https://www.clarin.eu/>
- iLOD Cloud
Resources from the LOD cloud available as IPFS
- LRE-Map
<https://lremap.eira.info/>
- META-SHARE
<http://www.meta-share.org/>
- OLAC
<http://www.language-archives.org/>

Search

BROWSE

- All of Linghub
- Communities & Collections
- By Issue Date
- Authors
- Titles
- Subjects

MY ACCOUNT

- Login
- Register

DISCOVER

- Author
- Christian Chiarcos (379)
- Gilles Sérasset (50)
- Henry Kučera (16)
- W. Nelson Francis (16)

Copyright © 2020 All Rights Reserved by Prêt-à-LLOD Project.

Horizon 2020
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182.

Figure 4: Main page of Linghub

3.2.2. Resources display

An icon is displayed if the metadata from the resources are defined by using standard metadata vocabularies only (tick) or not (cross). This allows the user to identify data that follow standards.

Prêt-à-LLOD

Linghub Home / Test Community / testOlac / View Item

Show simple item record

Ugaritic literature : a comprehensive translation of the poetic and prose texts


This resource is described using Linked Data vocabularies standards recommended by Prêt-à-LLOD

dc.creator	Gordon, Cyrus Herzl, 1908-
dc.date	1949
dc.date.accessioned	2021-06-17T15:22:31Z
dc.date.available	2021-06-17T15:22:31Z

Figure 5: Automatically generated “tick” icon for resources using recommended Linked Data vocabularies, with message on mouse hover

3.2.3. URL verification

An icon is displayed to notify the user whether a URL resolves or not. The following fields are currently checked: dcat.accessURL, dcat.endpointURL, dcat.downloadURL, foaf.homepage and dcat.landingPage.

rdf.type	http://www.w3.org/ns/dcat#Distribution
dcat.accessURL	http://zhishi.me/sparql 
dcat.mediaType	api/sparql

URL may not be accessible

Figure 6: Automatically generated “cross” icon for urls that do not resolve, with message on mouse hover

3.2.4. Language mapping

An icon is displayed with a message on mouseover when the language of the resource can be mapped to the standard ISO format.


dc.identifier.uri	localhost:8080/xmlui/handle/123456789/202439
dc.language.iso	swe 
dc.rights	CC-BY <div data-bbox="438 1153 1236 1198" style="background-color: black; color: white; padding: 2px; font-size: x-small;">This field was automatically generated based on the metadata provided. It follows Prêt-à-LLOD standards recommendations</div>
dc.title	Nils Matsson Kiöping's journeys
dcterms.hasPart	https://annohub.linguistik.de/resource/HRDjLbMKwdSWyceEyYwbAQDFK9gyHv6aXij/sWpY18Q=/file/cMAjvnjG2MuEBQdCUE6J+Ftmj1/sKZ3jOwSF8y7zOmg=
dcterms.language	http://lexvo.org/id/iso639-3/swe
dcterms.type	http://www.resourcebook.eu/lremap/owl/lremap_resource.owl#Corpus

Figure 7: Automatically generated dc.language.iso field for recognised language, with message on mouse hover

3.2.5. ODRL Policy mapping

An icon is displayed with a message on mouseover when the license/policy of the resource can be mapped to the ODRL policies.

dc.rights	http://creativecommons.org/licenses/by-nc-sa/3.0/ ✓	Login Register
dc.rights	Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) ✓	
dc.subject	bridging anaphora	
dc.subject	textual coreference	
dc.title	Extended Textual Coreference and Bridging Relations in PDT 2.0	
dc.type	http://purl.org/dc/dcmitype/Text	
dc.type	http://www.language-archives.org/vocabulary/type#primary_text	
dc.type	corpus	
rdf.type	http://www.w3.org/2000/01/rdf-schema#Resource	
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-LLOD standards recommendations
dc.available	2012-02-20T13:56:58Z	

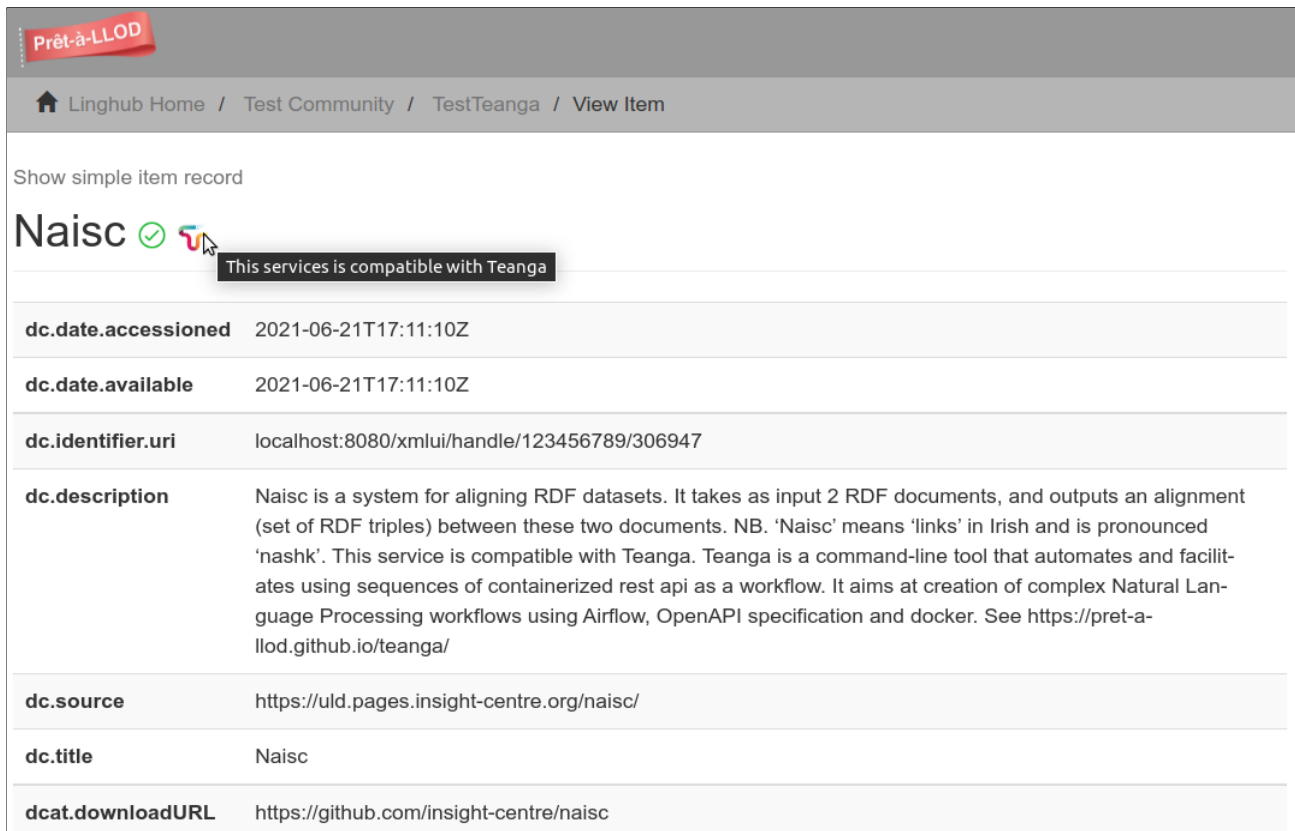
Figure 8: Automatically generated odrl.Policy field for recognised policy, with message on mouse hover

dc.identifier.uri	localhost:8080/xmlui/handle/123456789/202439	
dc.language.iso	swe ✓	
dc.rights	CC-BY ✗	
dc.title	Nils Matsson Kjöping's journeys	This field does not follow Prêt-à-LLOD standard recommendations

Figure 9: Display of a “cross” icon for policies non compatible to ODRL, with message on mouse hover

3.2.6. Teanga service identification


An icon is added to a resource (near its title) in the UI to mark if the resource is a service compatible with Teanga.



Prêt-à-LLOD

Linghub Home / Test Community / TestTeanga / View Item

Show simple item record

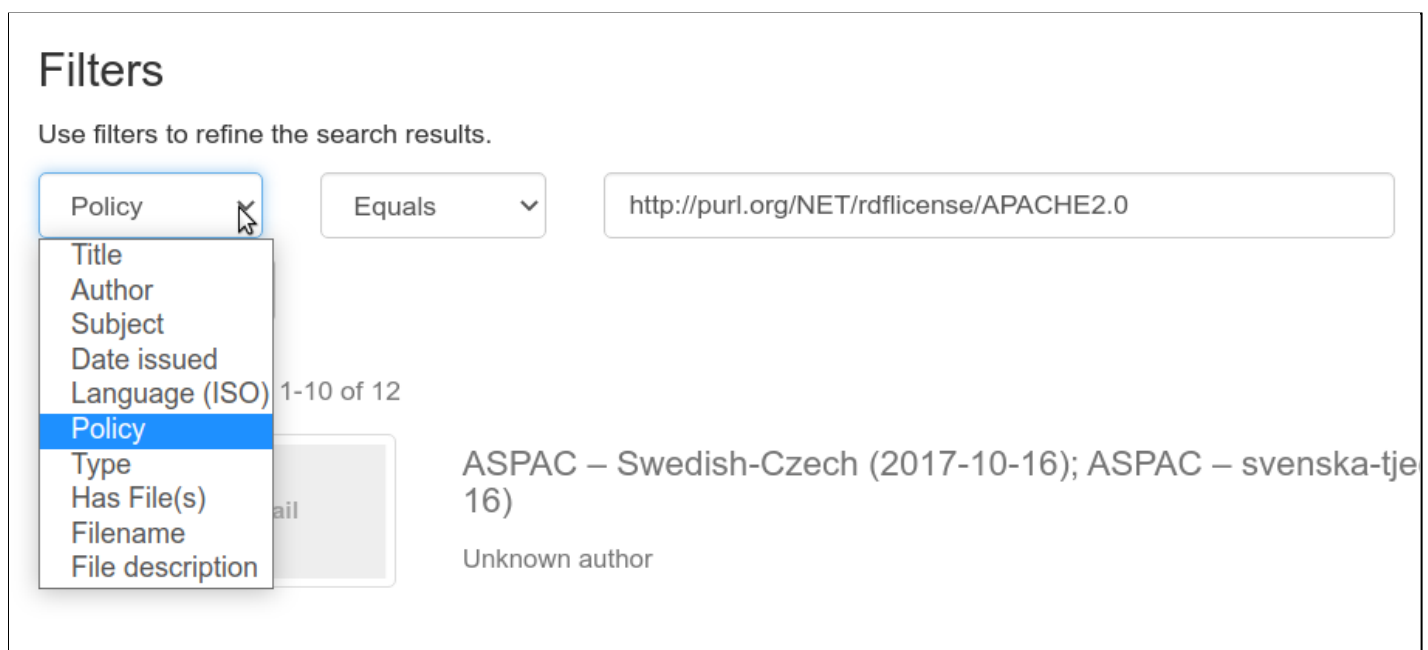
Naisc  This services is compatible with Teanga

dc.date.accessioned	2021-06-21T17:11:10Z
dc.date.available	2021-06-21T17:11:10Z
dc.identifier.uri	localhost:8080/xmlui/handle/123456789/306947
dc.description	Naisc is a system for aligning RDF datasets. It takes as input 2 RDF documents, and outputs an alignment (set of RDF triples) between these two documents. NB. 'Naisc' means 'links' in Irish and is pronounced 'nashk'. This service is compatible with Teanga. Teanga is a command-line tool that automates and facilitates using sequences of containerized rest api as a workflow. It aims at creation of complex Natural Language Processing workflows using Airflow, OpenAPI specification and docker. See https://pret-a-llod.github.io/teanga/
dc.source	https://uld.pages.insight-centre.org/naisc/
dc.title	Naisc
dcat.downloadURL	https://github.com/insight-centre/naisc

Figure 10: Display of a Teanga logo for Teanga-compatible services, with message on mouse hover

3.2.7. Advanced search

The fields “language”, “policy” and “type” are added to the DSpace standard list of filters to allow the user to search for datasets also on these fields.



Filters

Use filters to refine the search results.

Policy Equals

- Title
- Author
- Subject
- Date issued
- Language (ISO) 1-10 of 12
- Policy**
- Type
- Has File(s)
- Filename
- File description

ASPAC – Swedish-Czech (2017-10-16); ASPAC – svenska-tje 16)

Unknown author

Figure 11: Filters “Language (ISO), “Policy” and “Type” added to the standards filters provided initially

Language (ISO)	Policy	Type
eng (1308)	http://purl.org/NET/rdflicense/APACHE2.0 (21)	Text (3882)
deu (1009)	http://purl.org/NET/rdflicense/allrights reserved (12)	Sound (2076)
fra (890)	http://purl.org/NET/rdflicense/cc-by4.0 (3)	archives sonores (1290)
spa (611)		

Figure 12: Discovery (browsing) fields for “Language (ISO), “Policy” and “Type” added to the standards filters provided initially

3.3. ODRL API and query form

3.3.1. API

Overview. An API to facilitate the operations related to licenses was designed, developed and deployed. The API operates over arbitrary policies represented in ODRL using the ODRL for Language Resources profile, although a dataset of commonly used licenses was also provided. An overview of the data model, data and services is shown in Figure 13, which is the Policy Driven Data Management (PDDM) available under a specific URL²⁵.

²⁵ <https://pddm-pal.oeg.fi.upm.es/>

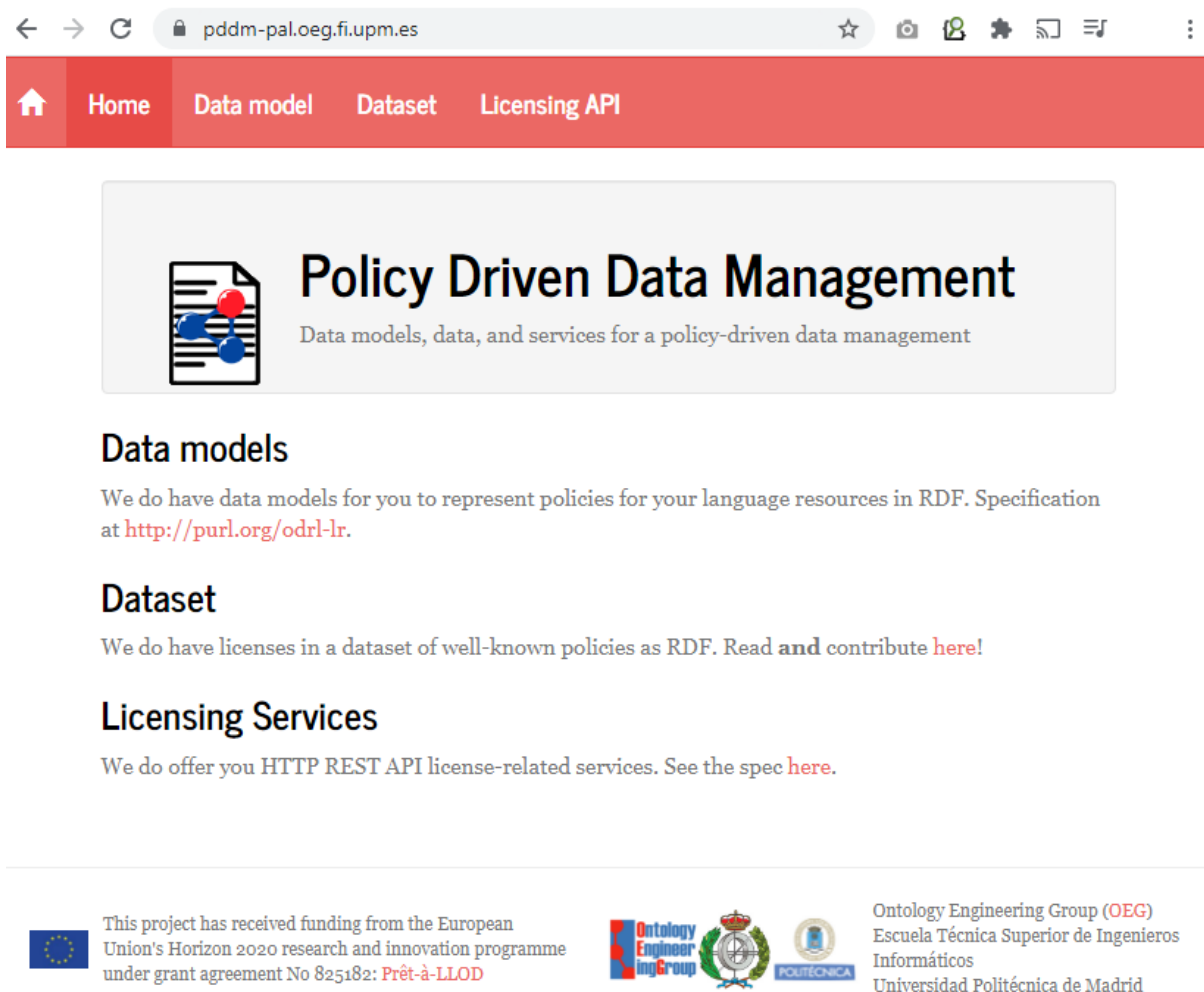


Figure 13: Policy Driven Data Management portal

Data model. An initial (and deprecated) version of the data model is served under a specific URL²⁶. This data model is an ODRL Profile, fully compliant with the mechanisms of profile creation foreseen by the official specification. At the time of edition of this document, the data model is not in its final form, as discussions within the W3C ODRL Community Group are pending. The latest version is available online²⁷. In order to gather the most realistic requirements, cooperation was established with the Institute of Language and Speech Processing (a division of Athena Research Center), as maintainers of the CLARIN Greece node and relevant actors of the licenses in the European Language Grid.

²⁶ <http://purl.org/odrl-lr>

²⁷ <https://rdflicense.linkeddata.es/profile.html>

W3C Community Group Draft Report

ODRL Profile for Policies of Language Resources and Technologies

Draft Community Group Report 27 August 2021

Latest published version:
<https://www.w3.org/TR/odrl-ir/>

Latest editor's draft:
<https://w3c.github.io/odrl/>

Editors:
 Penny Labropoulou (ILSP, Athena Research Center)
 Victor Rodriguez-Doncel (Universidad Politécnica de Madrid)

Participate:
[GitHub w3c/odrl](#)
[File an issue](#)
[Commit history](#)
[Pull requests](#)

Copyright © 2021 the Contributors to the ODRL Profile for Policies of Language Resources and Technologies Specification, published by the ODRL Community Group under the [W3C Community Contributor License Agreement \(CLA\)](#). A human-readable [summary](#) is available.

Abstract

This document presents the ODRL Profile for accessing Language Resources and Technologies (LRTs). It aims to support entities (repositories, infrastructures, archives, libraries, etc.) that enable the sharing of

Figure 14: ODRL Profile for Policies of Language Resources and Technologies

Data. A first collection of licenses was gathered (corresponding to the most popular used, such as the Creative Commons, or Apache licenses) and published in an open git repository of the Prêt-à-LLOD project²⁸. In order to gain in terms of sustainability, these licenses were transferred to another repository managed by the W3C ODRL Community Group, with more long term life expectancy²⁹.

API. An HTTP REST API was developed and deployed in the Policy Driven Data Management portal. This API has been documented using Open API standards (Swagger), and is available online³⁰ as can be seen in Figure 15.

²⁸ <https://github.com/Pret-a-LLOD/pddm/tree/master/data/licenses>

²⁹ <https://github.com/w3c/odrl/tree/master/bp/license>

³⁰ <https://pddm-pal.oeg.fi.upm.es/swagger-ui.html>

Pret-a-LLOD Policy Driven Data Management (PDDM) REST API 1.0.0

[Base URL: `pddm-pal.oeg.fi.upm.es/`]
<https://pddm-pal.oeg.fi.upm.es/v2/api-docs>

This is the documentation for the HTTP REST API for the Policy Driven Data Management. This HTTP REST API based on [JODRLAPI](#). You may want to see a sample Javascript [client](#) in in this link (see the source code).

[Apache 2.0](#)

Authorization Authorization Controller

- POST** `/authorizedS` Queries if the license is can be used

Compatibility Compatibilities Controller

- GET** `/compatibilityPair` Verifies the compatibility of two licenses
- GET** `/minimumCompatibilityPairLicenses` Get the minimum requirements for the compatibility of two licenses

License License Controller

- POST** `/license` postLicense
- GET** `/license/` getLicenses
- GET** `/license/{id}` getLicense
- DELETE** `/license/{id}` deleteLicense
- GET** `/loadDefaultLicenses` loadDefaults


Figure 15: PDDM REST API

3.3.2. Query form

One of the goals of the current task is to help ensure the lawful use of data and language technologies. We approach this by making the information about licenses and policies understandable and easily accessible to non-specialists. Through the platform, we provide a form for the users to verify whether they can use the data or not. This was developed in collaboration with project partner UPM, where they developed the ODRL API further to meet the needs of this application.

The form consists of three parts. The first one is concerned with the purpose of the use of the dataset (research or commercial). The second one is related to the affiliation of the user, whether it represents a company, or an academic institution/individual. Finally, the last field of the form refers to the duration the dataset is needed (where the user picks a date). The form has a limited number of options at the moment, but a member of the UPM team is currently working on a research collaboration which will look deeper at the different dependencies and requirements of the policies, and will extend the options accordingly.


Note that we can only provide such a function for datasets for which we identified and mapped an ODRL policy from the metadata given. The form is shown on the main display page of the resource (see figure 16 below)


dcterms.rights	Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)
dcterms.rights	http://creativecommons.org/licenses/by-nc-sa/3.0/
dcterms.subject	bridging anaphora
dcterms.subject	textual coreference
dcterms.title	Extended Textual Coreference and Bridging Relations in PDT 2.0
dcterms.type	corpus
dcterms.type	Text
odrl.Policy	http://purl.org/NET/rdflicense/cc-by-nc-sa3.0 

Check resource access

Choose a purpose

Who are you?

From date 

To date 

Authorized	Yes
------------	-----

Figure 16: Form allowing the user to verify whether they can use a dataset or not

3.4. SPARQL endpoint

As mentioned in section 1, DSpace caters for Linked Data thanks to the integration of Jena Apache Fuseki³¹. A SPARQL endpoint to query the data has been created and is currently accessible on the address <http://140.203.155.44:8001/dataset.html?tab=query&ds=/dspace>. (see figure 17 to see the platform). A more user-friendly url will be used, and will redirect to the endpoint in the future. The data is currently being populated in the triple store.

³¹ <https://jena.apache.org/documentation/fuseki2/>

Apache Jena Fuseki

dataset manage datasets help

Server status: ●

Dataset: /dSPACE

query upload files edit info

SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

EXAMPLE QUERIES

Selection of triples Selection of classes

PREFIXES

rdf rdfs owl xsd

SPARQL ENDPOINT: /dSPACE/sparql

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```

1
2
3 SELECT ?subject ?predicate ?object
4 WHERE {
5   ?subject ?predicate ?object
6 }
7 LIMIT 25

```

QUERY RESULTS

Table Raw Response

Figure 17: SPARQL query form

4. Resources

We describe here all the resources currently available in the new Linghub platform. Note that resources are included in Linghub only if they contain at least a title in the metadata, as this information is a minimum quality requirement for the data.

4.1. CLARIN

4.1.1. Overview

CLARIN ERIC is the European Research Infrastructure for Language Resources and Technology. CLARIN is a networked federation of language data repositories, service centres and centres of expertise. It makes digital Language Resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. Its digital infrastructure offers data, tools and services to support research based on language resources. Its data covers many repositories, such as the ZIM Centre for Information Modelling, Bavarian Archive for Speech Signals, and many more.

CLARIN is using its own metadata scheme, *cmdi*³², and makes the dump of their data available as xml files [online](#). The use of a specific scheme developed by CLARIN does not allow to easily harmonize with other repositories and resources. By including it in the Linghub platform and adapting the metadata to standards identified by the community, we make it accessible to communities without them having to learn yet another schema.

4.1.2. CLARIN in Linghub

At the writing of this deliverable, according to the [CLARIN Virtual Language Observatory](#), the repository contains 1,261,633 records. However the search through their platform yields 791,103 results instead.

Data from CLARIN was harvested for the first version of Linghub. We attempted to make a new harvest of the more recent data, however there have been changes in the metadata scheme and our harvester was not able to collect all the information from the *cmdi* format in this new harvest. Therefore at the moment we are using the first harvest used in the first version of Linghub, which contains 213,027 resources, and will work in the following months on updating the harvester to include the *cmdi* format changes. We identified through the language mapping module 49 languages covered by the repository.

4.2. OLAC

4.2.1. Overview

OLAC is “an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.” It was founded in 2000.

OLAC metadata is made available through [OAI-PMH](#), and also as a [nightly dump](#). It harvests language resources repositories with whom they established an agreement, such as the ELRA catalogue³³. ELRA is a non-profit-making organisation in the domain of Language Resources and Human Language Technologies, promoting access to such resources available in a great number of languages.

OLAC also developed their own metadata set and controlled vocabularies³⁴ to describe language resources. It is based on the complete set of Dublin Core metadata terms DCMT, but the format allows for the use of extensions to express community-specific qualifiers. It is often contrasted to IMDI (ISLE Metadata Initiative). A selection of recommendations and standards are also available on their website, such as the standards OLAC archives must follow in implementing a metadata repository³⁵ and best practice recommendations for language resource description.³⁶

³² <http://www.clarin.eu/cmd/1>

³³ <http://catalogue.elra.info/en-us/>

³⁴ <http://www.language-archives.org/OLAC/metadata.html>

³⁵ <http://www.language-archives.org/OLAC/repositories.html>

³⁶ <http://www.language-archives.org/REC/bpr.html>



4.2.2. OLAC in Linghub

The full metadata set of Linghub resources was collected through OAI-PMH on the 15/07/2021, as the nightly dump has proven to be a broken XML file and therefore unusable. 645 different languages were identified by our language mapping tool among the 442,501 resources extracted.

4.3. old.datahub

4.3.1. Overview

Old.datahub.io is the previous version of the current *datahub.io* platform, before the founding organization, Open Knowledge International³⁷, changed direction with a completely new architecture and setup, and decided to focus on data (as opposed to metadata) publication, with a selection of datasets for free and a Premium Data Service for additional or customised data. As a result, some datasets from old.datahub.io were not migrated to the new platform (because many of them were metadata only and did not come with the data itself). This results in a much lower number of resources compared to what was made available in the original platform. As in Linghub we are concerned with access to metadata (and links to the data within the metadata), we decided to harvest resources from the old.datahub platform, which extracts metadata from 981 different organisations.

4.3.2. old.datahub in Linghub

The data imported in the datahub collection in Linghub has been harvested using the old.datahub API in June 2021. This collection contains 2,615 resources.

4.4. iLOD

The iLOD dataset was created as part of task 5.3. It contains data from the LOD cloud, and makes it available through a peer-to-peer decentralized storage infrastructure using the InterPlanetary File System (IPFS). A more extensive description of this dataset, how it was created and the technologies involved, are presented in section 5. It contains 90,598 resources at the date of writing this deliverable, and will be maintained in Linghub with new data added using the REST API functionality of Linghub.

³⁷ <https://okfn.org/>



4.5. LRE Map

4.5.1. Overview

The LRE Map platform, initiated by ELRA and FlareNet at LREC 2010, collects information on both existing and newly-created language resources during the submission process of the LREC conference.

4.5.2. LRE Map in Linghub

We attempted to harvest more recent data from the LRE-Map website, but the source code of the website is non standard and has changed since the last harvest in the first version of Linghub, and it proved not to be possible to crawl the platform. We are therefore using the data harvested from the first version of Linghub in 2014, which contains 1,455 resources.

4.6. META-SHARE

4.6.1. Overview

META-SHARE is an open network of repositories for sharing and exchanging language data, tools and related web services. META-SHARE has developed its own metadata model for its own purpose, that is to serve the needs and requirements of the open distributed facility for sharing and exchanging resources of META-NET. META-SHARE has its own structured way of organizing resources. Since the aim of Linghub is to provide discovery and access to resources by the means of standard metadata vocabularies (such as Dublin Core, dcat), we are using a conversion to the Linghub standard metadata schemas.

4.6.2. META-SHARE in Linghub

The META-SHARE data available in Linghub was provided to us by the European Language Resource Association (ELRA). The anonymised dump of the data in the META-SHARE xml format was extracted on 17/06/2021, with 5,775 resources that we converted and integrated into Linghub, covering 160 different languages that we identified using our language mapping module.

4.7. Annohub

The Annohub dataset (Abromeit et al., 2020) was created and provided by project partner GUF. It is an open license dataset containing metadata about annotation and language information harvested from annotated language resources like corpora freely available on the internet. The dataset was created using a workflow that automatically generates metadata and provides subsequent curation of the results by domain experts. The generated metadata which is integrated to existing resources includes information about syntax and morphology, annotation models, language information (found as a tag or detected by the language identification tool), formalised concepts from the Ontologies of Linguistic Annotations (OLiA) corresponding to the detected annotations. In the future, this workflow will serve Linghub by



augmenting its resources with missing information, such as language. The current dataset created from the workflow is made up of 615 resources and covers 2,760 languages.

4.8. Teanga services

Services compatible with Teanga³⁸, the workflow aimed at easily using and combining NLP services together as a pipeline, are imported in Linghub in a separate collection. The initial set comprises 47 services, among which Naisc³⁹ and all the different services made available from DKpro⁴⁰, providing ready-to-use software components for natural language processing. These services will be constantly updated as new compatible services will be made available throughout the course of the project by the project partners, and in the future by other collaborators. Newly converted services will be added to the platform progressively using the REST API.

5. Linked Data Sustainability through IPFS

The proliferation of the World Wide Web and the Semantic Web applications has led to an increase in distributed services and datasets. This increase has put an infrastructural load in terms of availability, immutability, and security, and these challenges are being failed by the Linked Open Data (LOD) cloud due to the brittleness of its decentralisation. To address this, we have developed the iLOD system and published a dataset that is using the IPFS technology. IPFS uses content-based addressing, instead of location-based addressing, which makes the application's physical location transparent and ensures the content remains unique through all nodes.

iLOD is a dataset sharing system that leverages content-based addressing to support a resilient internet, and can speed up the web by getting nearby copies. We pre-processed Laundromat LOD (Beek et al., 2014) containing approximately 0.2 million datasets. iLOD capitalizes the Header, Dictionary, Triples (HDT) format and the IPFS technology to ensure data preservation by storing datasets securely across multiple locations. After the pre-processing (involving cleaning and dropping very small datasets having less than 1000 triples) and the conversion of datasets into HDT format, more than 90,000 datasets were added to iLOD together with their metadata information. The metadata contains links to other datasets (extracted using a linking algorithm), and statistics of a dataset like total number of tuples or cluster information. The linking algorithm works on partial matching of the object of the triple of one dataset to the subject of the triple of the other dataset. The dataset linking algorithm found 719,253 links between these datasets and around 32% datasets are linked to at least ten or more other datasets. By finding links between different datasets, connected components can be easily computed and one connected component can be considered as the cluster. This is the cluster information stored in the metadata. Our dataset clustering algorithm identified 12,324 clusters with 67% clusters having at least six or more datasets. iLOD is connected with Dspace and shares the metadata information. The DSpace API enables the data sharing of new data between iLOD and Dspace.

³⁸ <https://teanga.io/>

³⁹ <https://github.com/insight-centre/naisc>

⁴⁰ <https://dkpro.github.io/>



6. Maintenance

Task T5.3 has been completed with the deliverable of the platform. However the maintenance of Linghub through backup, update of existing resources, adding new resources, upgrades, etc. is a long term and continuous task. Some more developments in this regard are therefore expected and will be reported separately in the following months.

7. Conclusions

This deliverable reports on the development of the Linghub platform, the language resource discovery portal. All the main targets of availability, quality and content monitoring, accessibility and repository management, long-term storage, access to the data and related services, and engaging with the language resources communities have been met through the delivery of the Linghub platform.

8. Technical Documentation

We are maintaining an internal technical documentation of Linghub, with all details about the settings, customizations, installation of the main platform and all its components, explanations on how to restart the platform, and some useful links to the DSpace documentation. This is a document evolving constantly.

<https://docs.google.com/document/d/1p1rPjYy-QFz3a6FM7RQQ6Sy0SZ9fcjFSE3vXlbSjMmk>

9. References

Abromeit F., Fäth C., Glaser L., 2020, “Annohub – Annotation Metadata for Linked Data Applications”, In *proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, online⁴¹

W. Beek, L. Rietveld, H. R. Bazoobandi, J. Wielemaker, S. Schlobach, Lod Laundromat: A uniform way of publishing other people’s dirty data, *The Semantic Web – ISWC 340 2014*, Springer International Publishing, Cham, 2014, pp. 213–228

⁴¹ <https://aclanthology.org/2020.ldl-1.6/>

