



**HAL**  
open science

## D2.6: Ontology of licencing, ownership and conditions of use (V1.0)

Jason Carvalho, Enrico Daga, Marco Gurrieri, Andrea Scharnhorst

### ► To cite this version:

Jason Carvalho, Enrico Daga, Marco Gurrieri, Andrea Scharnhorst. D2.6: Ontology of licencing, ownership and conditions of use (V1.0). Open University. 2023. halshs-04265780

**HAL Id: halshs-04265780**

**<https://shs.hal.science/halshs-04265780v1>**

Submitted on 31 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



*Polifonia: a digital harmoniser for musical heritage knowledge, H2020*

## **D2.6: Ontology of licencing, ownership and conditions of use (V1.0)**

<b>Deliverable information</b>	
WP	WP2 Ontology Network
Deliverable dissemination level	PU Public
Deliverable type	R Document, report
Lead beneficiary	OU
Contributors	KNAW, CNRS, KCL, CNAM, UNIBO, NUIG
Document status	Final
Document version	V1.0
Date	October 30, 2023
Authors	Enrico Daga (OU), Jason Carvalho (OU), Marco Gurrieri (CNRS), and Andrea Scharnhorst (KNAW)
Peer review	Jacopo De Berardinis (KCL), Raphaël Fournier-S'niehotta (CNAM)



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746*

PAGE INTENTIONALLY BLANK

## Project Information

**Project Start Date:** 1st January 2021  
**Project Duration:** 40 months  
**Project Website:** <https://polifonia-project.eu>

## Project Contacts

### Project Coordinator

**Valentina Presutti**

ALMA MATER STUDIORUM -  
UNIVERSITÀ DI BOLOGNA  
Department of Language, Literature and  
Modern Cultures (LILEC)

E-mail: [valentina.presutti@unibo.it](mailto:valentina.presutti@unibo.it)

### Project Manager

**Marta Clementi**

ALMA MATER STUDIORUM -  
UNIVERSITÀ DI BOLOGNA  
Research division

E-mail: [marta.clementi3@unibo.it](mailto:marta.clementi3@unibo.it)

## POLIFONIA Consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	OU	THE OPEN UNIVERSITY	United Kingdom
3	KCL	KING'S COLLEGE LONDON	United Kingdom
4	NUI GALWAY	NATIONAL UNIVERSITY OF IRELAND GALWAY	Ireland
5	MiC	MINISTERIO DELLA CULTURA	Italy
6	CNRS	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	France
	SORBONNE	SORBONNE UNIVERSITE (LinkedTP)	France
7	CNAM	CONSERVATOIRE NATIONAL DES ARTS ET METIERS	France
8	NISV	STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID	Netherlands
9	KNAW	KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETEN- SCHAPPEN	Netherlands
10	DP	DIGITAL PATHS	Italy

## Project Summary

European musical heritage is a dynamic historical flow of experiences, leaving heterogeneous traces that are difficult to capture, connect, access, interpret, and valorise. Computing technologies have the potential to shed a light on this wealth of resources by extracting, materialising and linking new knowledge from heterogeneous sources, hence revealing facts and experiences from hidden voices of the past. Polifonia makes this happen by building novel ways of inspecting, representing, and interacting with digital content. Memory institutions, scholars, and citizens will be able to navigate, explore, and discover multiple perspectives and stories about European Musical Heritage.

Polifonia focuses on European Musical Heritage, intended as musical contents and artefacts - or music objects - (tunes, scores, melodies, notations, etc.) along with relevant knowledge about them such as: their links to tangible objects (theatres, conservatoires, churches, etc.), their cultural and historical contexts, opinions and stories told by people having diverse social and artistic roles (scholars, writers, students, intellectuals, musicians, politicians, journalists, etc), and facts expressed in different styles and disciplines (memoire, reportage, news, biographies, reviews), different languages (English, Italian, French, Spanish, and German), and across centuries.

The overall goal of the project is to realise an ecosystem of computational methods and tools supporting discovery, extraction, encoding, interlinking, classification, exploration of, and access to, musical heritage knowledge on the Web. An equally important objective is to demonstrate that these tools improve the state of the art of Social Science and Humanities (SSH) methodologies. Hence their development is guided by, and continuously intertwined with, experiments and validations performed in real-world settings, identified by musical heritage stakeholders (both belonging to the Consortium and external supporters) such as cultural institutes and collection owners, historians of music, anthropologists and ethnomusicologists, linguists, etc.

## Executive Summary

In research workflows under the paradigm of Open Science (standing for reproducibility of research, open access to knowledge, and societal responsibility of research) licences play an increasing role. With digitisation and automatic information processing, licences become important to also to guide the actions of machines, for example, in supporting the exploration and selection of resources and auditing their fair reuse. In the context of Polifonia we deal primarily with licences which come with content provided in the public sphere by cultural heritage institutions. But, we are also dealing with other source material: for instance information scrapped from websites, and we produce and re-use software which also comes with a licence, such as the resources catalogued by the [musoW registry](#) of musical resources on the Web.

There are various issues when it comes to licences: - there is a large variety of licences and copyright statements used in the domain of musical content - the information about licences is not always added to metadata or not added in a standardised way, but often 'hidden' in plain text on websites - licences regulating the access to and use of a webservices (e.g., repositories) and licences regulating the access and use of content provided via webservices (e.g. datasets in a repository) are kind of entangled - there might be various, sometimes contradicting each other, licence information available for a certain data collection.

In this deliverable, we focus on the problem of extracting licence information from Web resources. More specifically, we look into the coverage of licence metadata in data registries, such as musoW a catalogue in which all main data components used by Polifonia are registered, next to a large number other sources. We set up pipelines to check for licence information, and where possible to enrich it, text-mining the original websites/sources to which the catalogue refers. We do so with the aid of Large Language Models (LLM). LLMs are receiving increasing attention in numerous applications, including knowledge extraction, but little work has been done so far in extracting and linking licence information with help of them. Working with semantic web principles as our core technology means, we are in particular devoted to design workflows where licence information can be turned into structured data (best expressed as so-called semantic artefacts); expressed in form of ontologies and knowledge graphs. As a result, we develop iterative workflows where LLM use is combined with querying structured information as coded in ontologies and knowledge graphs.

We depart from the source material we use in Polifonia and start with an overview about the Polifonia datasets with the aim to define our problem space (Chapter 2). We devote an entire chapter to discuss related work (3) through which we render the possible solution space. Here, we briefly summarise the current discourse among those who further detail rules for FAIR implementation (Section 3.1); we describe the current state of art if it comes to the knowledge representation for licences and terms of use (Section 3.2). We give an overview of the prominent approaches to licence expressions on the Web: MPEG, CC-REL, and ODRL. The latter we further address in Section 3.3. We further touch upon the problem of reasoning with licences on the Web (Section 3.4). We close this chapter with a description of the use of Large Language Models - as we will apply them later on in our workflows (3.5). Chapter 4 concerns a specific workflow how to extract licence information from web resources with a Large Language Model (LLM). Here, we concentrate on the *musoW* resource which entails many relevant resources including all Polifonia data components which are also registered in the Polifonia Research Ecosystem. The workflow leads to an enrichment of the original licence information available in the *musoW* catalogue. Chapter 5 deals with the task of Knowledge Graph Construction. It entails our design to formalise the extracted information and align it to existing knowledge graphs, and in particular how to best integrate the data into a *Licences Knowledge Graph*, a core output of this deliverable. As one result we also gain a deeper insights in which licences are used *in the wild*, meaning in the practices of many musicologists and music documentalists. Therefore, the chapter is followed by an Evaluation (6), complemented by the Polifonia Fair Section (7)- the check list if we oblige to the FAIR guides agreed upon in Polifonia. We conclude the deliverable with Chapter 8.

The implications of the work done in this deliverable (D2.6), for the further development of the Polifonia Ecosystem, in particular, how these results influence how we treat licences in the Polifonia Research Ecosystem (and the related framework), will be discussed and reported in the Final Data Management Plan (D7.3).

## Document History

Version	Release date	Summary of changes	Author(s) - Institution
v0.1	20/04/2023	Deliverable layout	Enrico Daga (OU)
v0.2	10/06/2023	Deliverable structured and subsections	Enrico Daga (OU)
v0.3	25/08/2023	Introduction	Enrico Daga (OU)
v0.4	20/09/2023	Related work	Enrico Daga (OU)
v0.5	01/10/2023	Requirements analysis, executive summary, and related work	Marco Guerrieri (CNRS) and Enrico Daga (OU) and Andrea Scharnhorst (KNAW)
v0.6	10/10/2023	Executive summary, knowledge graph construction, and FAIR section	Enrico Daga (OU) and Andrea Scharnhorst (KNAW)
v0.7	11/10/2023	Experiments section completed and conclusions	Enrico Daga (OU) and Andrea Scharnhorst (KNAW)
v0.8	11/10/2023	Version sent to reviewers for feedback	Enrico Daga (OU) and Marco Guerrieri (CNRS) and Andrea Scharnhorst (KNAW)
v0.8r	20/10/2023	Received feedback from reviewers	Jacopo De Berardinis (KCL), Raphaël Fournier-S'niehotta (CNAM)
v0.9	30/10/2023	Release submitted to the coordinator.	Andrea Scharnhorst (KNAW), Enrico Daga (OU), Marco Guerrieri (CNRS)
v1.0	Final version	Submission to EU	Coordinator

# Table of contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Overview of Polifonia requirements (defining the problem space)</b>	<b>3</b>
2.1	Scenario . . . . .	3
2.2	Requirements from the survey / socio-technical map . . . . .	3
2.3	The musoW catalogue . . . . .	5
<b>3</b>	<b>Related Work (defining the solution space)</b>	<b>8</b>
3.1	FAIR principles in the context of computational licences and terms of use . . . . .	8
3.2	Knowledge representation for licences and terms of use . . . . .	9
3.3	Open Digital Rights Language (ODRL) . . . . .	10
3.4	Reasoning with licences on the Web . . . . .	12
3.5	Large Language Models for Knowledge Extraction . . . . .	13
<b>4</b>	<b>Extracting licence information from web resources with a Large Language Model (LLM)</b>	<b>14</b>
4.1	Methodology . . . . .	14
4.2	Experiments . . . . .	16
4.2.1	Data preparation . . . . .	16
4.2.2	Task 1: finding links in web pages . . . . .	17
4.2.3	Task 2: extract copyright, licence, and terms of use information from content . . . . .	17
4.2.4	Task 3: link licence descriptions to the licences database . . . . .	20
4.3	Results and discussion . . . . .	21
<b>5</b>	<b>Knowledge Graph Construction</b>	<b>25</b>
<b>6</b>	<b>Querying the Polifonia Licences Knowledge Graph</b>	<b>28</b>
<b>7</b>	<b>Compliance to the FAIR principles</b>	<b>32</b>
<b>8</b>	<b>Conclusions</b>	<b>34</b>



## 1 Introduction

Cultural heritage (CH) assets are guarded in the rich but also very scattered landscape of cultural heritage institutions. Increasingly, information documented in those assets is seen as research data [1]. Polifonia is located in the innovative forefront of working with CH data, and its researchers manage, further curate, enrich and analyse these assets together with their original data owner or guardians. Data-reuse inside of the Polifonia project appears typically through different activities and methods, spanning from downloading/copying documents or audio files to consuming their content (reading scores or librettos, listening to a song on a radio, ...).

For Polifonia researchers as well as the project stakeholders, legal compliance in the reuse of musical cultural heritage is vital. Work package 2 applies ontology engineering practices to model concepts, relations and entities in musical heritage. This approach opens up opportunities for answering key questions regarding the legal dimensions of reuse in collaborative research and innovation projects.

In this deliverable, we focus on the problems related to exploiting and extending music collection metadata to answer questions regarding the rights and terms of use of the data to support stakeholders such as music researchers, computational musicologists, and industry practitioners in selecting and evaluating the suitability of musical data sources for their projects. The objective is to assist users in understanding the legal requirements behind the reuse of assets with specific intellectual property rights.

Building on the background of managing licences and provenance on the Web as developed in the domain of library and information sciences, in Polifonia we apply state-of-the-art methods for the representation of licences and terms and conditions applicable to the variety of objects<sup>1</sup>.

We use a Knowledge Graph (KG) approach to integrate and query the metadata of datasets alongside licencing information. As result, we produce a Knowledge Graph of licence metadata to link the datasets and resources associated with project outputs with a computational representation of their terms of use, therefore making them ready for supporting recent advances in policy reasoning [2], licence compatibility [3] and policy propagation [4].

Our methodology is based on a qualitative assessment of the landscape of musical datasets used in the Polifonia project, performed by expert musicologists. This expert assessment uncovers a set of key requirements that we express in terms of *Competency Questions* (CQ), following good practices in ontology engineering.

Next, we analyse the landscape of representation and reasoning on computational licences and select the W3C Open Digital Rights Language (ODRL) as the reference representation method for computational licences. To support querying and reasoning over terms and conditions of licences, we select the DALICC collection of licences expressed as RDF/ODRL [5], which we reuse in our work.

Our starting point is the musoW catalogue [6], a high-quality metadata collection of musical resources on the Web. We link the musoW metadata related to licence information to the DALICC resources. One crucial aim is to extend the coverage of the musoW metadata by including licence information that may have been not provided by the original contributors of information to the musoW catalogue. In particular, we study how to apply a Large Language Model (LLM) for supporting the task of extracting and later linking of licencing information from web resources. The resulting extended metadata forms the basis of the Polifonia Licences Knowledge Graph.

In short, first we apply existing generic formalisations of licence information to enrich the metadata about licence. We use a central component in Polifonia to execute this - the musoW catalogue. Second, we cross-check of those for

Finally, we validate the suitability of terms of use expressed in ODRL to answer queries relevant to musicology researchers and, at the same time, the suitability of RDF licence catalogues such as DALICC to cover the breath of resources relevant to musicological research, considering the musoW catalogue as a reference. We demonstrate

<sup>1</sup>In what follows, we use the British spelling *licence* in the writing but keep the original form whenever referring to project or software names, as well as ontology terms (e.g. the Schema.org term `schema:license`).

how it is possible to answer key questions in relation to rights and permissions associated with datasets by applying a collection of licences represented with ODRL.

The rest of the deliverable is structured as follows. Chapter 2 first presents the research landscape of the Polifonia project relating the pilot applications with their data requirements (Section 2.2). Such requirements are expressed as competency questions (CQs) following good practices in ontology engineering. In Section 2.3, we report on the musoW catalogue and provide statistics regarding licence metadata. We observe how many resources don't have an explicit licence annotation and wonder whether we could improve licence annotations by leveraging Large Language Models (LLM). In Chapter 3 we discuss relevant literature considering FAIR principles in open data sharing (3.1), foundations of knowledge representation applied to licencing and terms of use (3.2), the Open Digital Rights Language (ODRL), which we reuse in our work (3.3), and how it can be applied to query and reason on licencing information on the web (3.4). We conclude the related work section with a brief overview of current applications of Large Language Models (LLM) for extracting knowledge from texts.

Chapter 4 reports on our experiments with applying an LLM to support extracting and linking licencing information from web resources. The method is evaluated for its capacity to automatically find relevant links, extract information about copyright and licencing, and link licence descriptions to the DALICC catalogue of computational licences. We apply our method to acquire licence information and extend the coverage of the musoW catalogue.

Next, we describe the pipeline to build the Polifonia Licences Knowledge Graph, combining data from musoW and DALICC (Chapter 5). In Chapter 6, we evaluate the KG by encoding the CQs designed in Section 2.2 into SPARQL query, and provide statistics of the musoW dataset from the point of view of the terms and conditions associated with the resources. We discuss how our work satisfies the FAIR principles for open science in Chapter 7 Finally, in Chapter 8 we discuss our contribution and sketch directions for future work.

## 2 Overview of Polifonia requirements (defining the problem space)

In this chapter, we observe the landscape of data requirements of Polifonia in order to illustrate the needs of music information practitioners. We look at this issue from three perspectives. First, following the Polifonia ontology engineering methodology, we design a Story focused on a researcher in computational musicology that wants to reuse an existing dataset. Next, we report on observations from a survey conducted in D1.1, where information needs of the pilot case studies were analysed also from the perspective of licencing and terms of use. Finally, we survey the musoW catalogue of musical resources on the Web [6], a key resource in Polifonia, whose metadata collection also include resources reused by the pilots and produced by the various work packages of Polifonia. Therefore, we consider the musoW metadata set the starting point for answering to those needs.

### 2.1 Scenario

For researchers and computational music practitioners, awareness about the licencing and terms of use of datasets and resources involved in the project is a crucial issue. This need can be illustrated in the following scenario based on the Sethus Persona and focused on the Tonalities pilot:

*Sethus wants to reuse the MEI scores encoded in the framework of the Josquin Research Project from the Stanford University<sup>1</sup>. This choice is grounded on the centrality of Josquin Des Près and on the high relevance of his musical output that perfectly responds to Sethus' analytical needs and purposes. For doing this, he decides to do a preliminary research about the licence covering the work of the Josquin Research Project and to contact the editors of the project in order to obtain their availability to share and to reuse their musical scores. Accordingly to the info collected on the project website and its GitHub repository, Sethus, who meanwhile received positive feedback from his colleagues in Stanford, realises that the MEI scores of the Josquin Research Project are covered by a CC-BY-SA 4.0 licence: he can therefore reuse the MEI scores (even for commercial purposes) by quoting the source of their origin, and modify whatever he wants on condition of alerting the reader about any change or modification. Sethus can now reuse these MEI files to enrich Tonalities' Online Collaborative Interface for Music Analytical Annotations [7], constituting a coherent corpus of music compositions that can be analysed using the theoretical models and the analytical categories and classes available in the interface. The MEI files are reused to display the scores on the screen in an interactive infrastructure that allows the user to select single notes, patterns and/or chords in order to apply a particular analytical category to the above mentioned selection, so creating an annotation that is registered by the system. All the validated annotations can be consulted, quoted, commented by other users, promoting the critical discussion and the methodological exchanges among analysts about their different viewpoints.*

Other Stories, among those ones that have been already published by the Pilots on the GitHub repository<sup>2</sup>, can suggest different scenarios concerning the issue of the publication rights and related requirements. With the sole exception of Access (whose pilot's goals do not deal with existing datasets), each pilot can articulate its specific needs and commitments towards licences and copyrights.

### 2.2 Requirements from the survey / socio-technical map

Here, we analyse the problem on a larger scale, looking into the breadth of pilot use cases of Polifonia.

The Socio-technical roadmap (ref D1.1) made emerge key requirements for the pilots in relation with their intention to reuse existing datasets. For this purpose a section of the Survey proposed by the WP1 has been dedicated to technical aspects, among which in particular the description of the external datasets elaborated in the framework

---

<sup>1</sup><https://josquin.stanford.edu>

<sup>2</sup><https://github.com/polifonia-project/stories>

of existing music and/or musicological projects. Each pilot has therefore been asked to characterise the existing data they want to reuse describing their origin, their nature (what standards), their format, as well as to check the possibility to assure an effective interoperability. The relevant section of the survey is repurposed here in Table 2.1.

The following table shows the part of the Survey where the pilots were asked to characterise the existing datasets they want to reuse. The sources and the datasets quoted in the next table focusing on licences and rights are referring to those ones identified in this first table.

**Table 2.1:** Excerpt of the survey conducted in D1.1 in relation to information needs of the pilot case studies.

<b>Legal issue</b>	<b>Under what licences are the sources provided?</b>	<b>Who has the rights to the used datasets?</b>
MUSICBO	Not known yet	Not known yet
TONALITIES	*CC BY-NC 3.0*CC BY-NC-SA*CC BY-SA*CC BY-SA 4.0*CPDL copyright licence*GNU GENERAL PUBLIC LICENSE, Version 3, 29 June 2007	Public domain datasets.
CHILD	We only consider books available in the public domain. All produced material is published free to reuse under a CC-BY licence.	Used sources are in the public domain. Produced data is OU copyright.
MEETUPS	We only consider books available in the public domain. All produced material is published free to reuse under a CC-BY licence.	Used sources are in the public domain. Produced data is OU copyright.
ACCESS	Not applicable (N/A)	Not applicable (N/A)
BELLS	*Ministry's Dataset are in CC-BY-SA.*Datasets not created by the ministry may be subject to copyright.	At the moment some sources are protected by copyright, but we are looking forward to have the permission for the application of CC licenses.
INTERLINK	*The dataset is licensed through CC-BY 4.0*Some parts (e.g. MySongBook transcripts in MIDI RDF) might be subject to copyright	Public domain datasets.
FACETS	Creative Commons licences.	Public domain datasets.
ORGANS	Not known yet	Dutch National Institute of Organ Art (NIvO).
TUNES	Not known yet	The Meertens Institute

Based on these preliminary results, the WP1 created a Spreadsheet document to report the requirements for the licencing and the terms of use for each of these datasets. The research methods have been based on a double-checking process of the data coming from musoW and from what declared in the on-line website of each listed external project (and in the GitHub repository of the project, if any). In particular this Spreadsheet document has been structured collecting in different columns: the name of the project, the URL of the website (if any), the declared licence(s), the presence of copyright, the access conditions (free or paying). A sublist of the reported licences have been created containing conditions and requirements for each licence with a focus on the case of reuse, modification and ShareAlike obligations.

From the scenario and requirements discussed so far, we can elicit the following competency questions (CQ):

- CQ1 Do you want to reuse an external dataset?
- CQ2 Does the licence covering the dataset permit to share?
- CQ3 If yes, what are the licence requirements for sharing the data?
- CQ4 Do you want to modify the existing dataset?
- CQ5 Does the licence covering the dataset permit to modify the original data?
- CQ6 If yes, what are the licence requirements for modifying the data?
- CQ7 Does the licence impose a ShareAlike obligation?
- CQ8 Does the licence impose any commercial restrictions?

## 2.3 The musoW catalogue

In this section we report on the current status of the musoW catalogue regarding licence metadata. The data reported is being observed on the 3rd October 2023 and results from querying the musoW SPARQL endpoint.

Listing 2.1 provides the SPARQL query summarising licence metadata, results are in Table 2.2. We also report the same statistics grouped by type: media (Table 2.3) – referring to audio/video databases; metadata (Table 2.4) referring to registries; and symbolic (Table 2.5) – referring to content encoded in symbolic means.

It can be seen that most resources don't have a specific licence associated with it (almost 70% of the registry). The reasons may vary:

- the resource does not have a specific licence;
- the information being not available at the time the metadata was curated, but now it is available;
- the curator overlooked the information.

The lack of sufficient licencing and terms of use information for published web resources is a well-known problem whose impact on the broader landscape of content reuse on the web cannot be underestimated. In the case of musoW catalogue, we are confident that most of the annotations are actually correct (or they were correct at the time of their retrieval). However, supporting curators in retrieving such information automatically without having to browse each one of the websites catalogued manually would certainly contribute to improving the quality and coverage of the musoW catalogue. Therefore, we dedicate Chapter 4 to experiment with Large Language Models (LLM) for improving licence annotations in musoW.

### Listing 2.1: SPARQL query to extract the list of values for the schema:licence metadata term in musoW

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <https://schema.org/>
SELECT ?obj ?label (count(?sub) as ?howmany)
WHERE {
  ?sub schema:license ?obj .
  optional { ?obj rdfs:label ?label}
}
group by ?obj ?label
order by desc(?howmany)
```

### Listing 2.2: SPARQL query to extract an overview of licences in the musoW catalogue for each type of resource (metadata, symbolic, or audio).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <https://schema.org/>

SELECT ?category ?obj ?label (count(?sub) as ?howmany)
WHERE {
  ?sub schema:category ?category .
  optional { ?sub schema:license ?obj . ?obj rdfs:label ?label}
}
group by ?category ?obj ?label
order by ?category desc(?howmany)
```

**Table 2.2:** Summary of licence annotations in the musoW dataset

https://w3id.org/musow/vocab/open-access	Open Access	274
https://w3id.org/musow/vocab/cc-by	CC-BY	50
https://w3id.org/musow/vocab/not-specified	Not specified	45
https://w3id.org/musow/vocab/copyright	Copyright	30
https://w3id.org/musow/vocab/unknown-licence	Unknown	29
https://w3id.org/musow/vocab/cc0	CC0	26
https://w3id.org/musow/vocab/cc-by-nc-sa	CC-BY-NC-SA	14
https://w3id.org/musow/vocab/cc-by-nc	CC-BY-NC	9
https://w3id.org/musow/vocab/gnu-gpl	GNU-GPL	5
https://w3id.org/musow/vocab/privative	Privative	4
https://w3id.org/musow/vocab/cc-by-sa	CC-BY-SA	3
https://w3id.org/musow/vocab/mit-license	MIT License	2
https://w3id.org/musow/vocab/cc-by-3.0	CC BY 3.0	2
https://w3id.org/musow/vocab/gnu-fdl	GNU-FDL	1
https://w3id.org/musow/vocab/cc-by-4.0	CC-BY 4.0	1
https://w3id.org/musow/vocab/cc-by-nc-nd	CC-BY-NC-ND	1

**Table 2.3:** Overview of licences in the musoW catalogue for resources of type "media".

https://w3id.org/musow/vocab/open-access	Open Access	48
https://w3id.org/musow/vocab/not-specified	Not specified	21
https://w3id.org/musow/vocab/cc-by	CC-BY	17
https://w3id.org/musow/vocab/copyright	Copyright	16
https://w3id.org/musow/vocab/unknown-licence	Unknown	12
https://w3id.org/musow/vocab/cc0	CC0	3
https://w3id.org/musow/vocab/gnu-gpl	GNU-GPL	3
https://w3id.org/musow/vocab/privative	Privative	3
https://w3id.org/musow/vocab/cc-by-nc-sa	CC-BY-NC-SA	2
https://w3id.org/musow/vocab/cc-by-nc	CC-BY-NC	2
-	-	2
https://w3id.org/musow/vocab/cc-by-4.0	CC-BY 4.0	1
https://w3id.org/musow/vocab/cc-by-sa	CC-BY-SA	1

**Table 2.4:** Overview of licences in the musoW catalogue for resources of type "metadata".

https://w3id.org/musow/vocab/open-access	Open Access	94
https://w3id.org/musow/vocab/cc-by	CC-BY	40
https://w3id.org/musow/vocab/not-specified	Not specified	23
https://w3id.org/musow/vocab/unknown-licence	Unknown	21
https://w3id.org/musow/vocab/copyright	Copyright	16
https://w3id.org/musow/vocab/cc0	CC0	9
https://w3id.org/musow/vocab/cc-by-nc	CC-BY-NC	5
https://w3id.org/musow/vocab/cc-by-nc-sa	CC-BY-NC-SA	4
-	-	4
https://w3id.org/musow/vocab/privative	Privative	3
https://w3id.org/musow/vocab/cc-by-nc-nd	CC-BY-NC-ND	1
https://w3id.org/musow/vocab/cc-by-3.0	CC BY 3.0	1
https://w3id.org/musow/vocab/gnu-gpl	GNU-GPL	1

**Table 2.5:** Overview of licences in the musoW catalogue for resources of type "symbolic".

<a href="https://w3id.org/musow/vocab/open-access">https://w3id.org/musow/vocab/open-access</a>	Open Access	176
<a href="https://w3id.org/musow/vocab/cc-by">https://w3id.org/musow/vocab/cc-by</a>	CC-BY	18
<a href="https://w3id.org/musow/vocab/not-specified">https://w3id.org/musow/vocab/not-specified</a>	Not specified	17
<a href="https://w3id.org/musow/vocab/cc0">https://w3id.org/musow/vocab/cc0</a>	CC0	15
<a href="https://w3id.org/musow/vocab/copyright">https://w3id.org/musow/vocab/copyright</a>	Copyright	14
<a href="https://w3id.org/musow/vocab/unknown-licence">https://w3id.org/musow/vocab/unknown-licence</a>	Unknown	9
<a href="https://w3id.org/musow/vocab/cc-by-nc-sa">https://w3id.org/musow/vocab/cc-by-nc-sa</a>	CC-BY-NC-SA	8
<a href="https://w3id.org/musow/vocab/cc-by-nc">https://w3id.org/musow/vocab/cc-by-nc</a>	CC-BY-NC	4
<a href="https://w3id.org/musow/vocab/gnu-gpl">https://w3id.org/musow/vocab/gnu-gpl</a>	GNU-GPL	3
<a href="https://w3id.org/musow/vocab/privative">https://w3id.org/musow/vocab/privative</a>	Privative	2
<a href="https://w3id.org/musow/vocab/cc-by-sa">https://w3id.org/musow/vocab/cc-by-sa</a>	CC-BY-SA	2
<a href="https://w3id.org/musow/vocab/mit-license">https://w3id.org/musow/vocab/mit-license</a>	MIT License	1
<a href="https://w3id.org/musow/vocab/cc-by-3.0">https://w3id.org/musow/vocab/cc-by-3.0</a>	CC BY 3.0	1
<a href="https://w3id.org/musow/vocab/gnu-fdl">https://w3id.org/musow/vocab/gnu-fdl</a>	GNU-FDL	1

### 3 Related Work (defining the solution space)

Recent advancements in music computing methods call for better supporting the understanding of the legal terms associated with music sources to computationally support the analysis of the legal issues surrounding the production of novel music from copyright-protected work that can be considered original work [8, 9]. Copyright information is included in many metadata standards (Dublin Core, Schema.org, and others)<sup>1</sup>. The application of knowledge representation techniques to associate rights to musical content is not new (for example, applying a Rights Expression Language such as MPEG-21 for supporting music distribution on the Web [10])

Policies on the Web means different things in relation to security and privacy [11, 12], access and control (ACL) [13, 14, 15], adaptable and context-aware systems (as a means to control the behaviour of complex systems [16]), and expression of legal knowledge (terms and conditions, licences) [17, 18]. In what follows we focus on the latter, considering approaches whose aim is to express and reason upon policies in the meaning of licences and *digital rights*, and limiting to the approaches designed to work with the WWW's architecture or principles.

In this chapter, we explore the computational treatment of licences from different perspective (FAIR data principles, knowledge representation and reasoning, and Web technologies). In addition, we give some background in the application of Large Language Models (LLM) for extracting knowledge, that we build upon in later chapters.

#### 3.1 FAIR principles in the context of computational licences and terms of use

In the discourse about FAIR and licences we can distinguish two streams: the fair use of licences (and here fair is used in a colloquial way of speaking) and the FAIR principles standing for Findable-Accessible-Interoperable and Re-usable) and their relation to licensing. Concerning the latter, licences most influence the FAIR aspect 'Re-usability'. The licence details what you are allowed with the data you retrieved. However, related to this is also the attribute of 'Access', as in repositories often the access preempts what you are supposed to do with the data.

The implementation of FAIR rules is closely linked to processes being machine-actionable. FAIR as such does not automatically imply that everything is *open*, even in the case of cultural heritage. There are various legitimate cases where access needs to be regulated. Think here in terms of 'sensitive data' if it concerns subjects, medical information and so on; or protecting cultural heritage (as still uncovered in soil) against robbery.

The natural, authoritative custodians for licences relevant for common goods is the Creative Commons Organisation<sup>2</sup>. So, for any recommendation for licence use one should depart from the set of licences they offer. Current discussion around FAIR implementation are not so much about which licence to apply, but how to regulate the access to a Digital Object (DO). As said above, one can regulate the access via access terms to repositories containing collections of Digital Objects. But, what about if DO's are circulating around and are re-used in various contexts? In such a case, access regulation should stay firmly with the DO, and one proposal is to encode it within a licence. This way ensuring that the access provision stands close to the Digital Object (DO), wherever this DO goes, and not to one of its original location.

Braukmann and Hugo – in the context of the Dutch ODISSEI project – worked on a proposal for more actionable licences for datasets under restricted access. They depart from the statement that "Open Access datasets typically come with creative commons licences that specify the conditions for reuse. For sensitive data, however, access conditions are often strict and much less standardised. More often than not the conditions under which access is granted are unclear to the researcher." (see [19]) They come up with a proposal how to change (extend or recombine) licences to also include special cases such as sensitive data.

<sup>1</sup>In this work, we don't discuss further methods to represent copyright information. Instead, we explore the realm of expressing terms of use that IP holders may grant to third parties.

<sup>2</sup><https://creativecommons.org/share-your-work/>



### 3.2 Knowledge representation for licences and terms of use

The legal domain has long been of interest to researchers in intelligent systems. Computer scientists developed theories and tools covering several research themes, spanning from the formalisation of legal concepts to the management of legal knowledge and the development of several research branches related to automated reasoning - rule processing, case-based reasoning, and deontic logic [20]. In knowledge engineering, legal ontologies emerged to support intelligent systems in tasks such as classification, knowledge sharing and decision support. The Semantic Web has inherited this interest, whose vision nicely fits crucial requirements of legal intelligent systems such as the strong accent on question answering rather than document retrieval [18]. Therefore, methodologies for the design of legal ontologies were developed in the context of the Semantic Web [21, 22, 23] as well as ontological frameworks like LRI-Core [24], the LKIF Core Ontology of Basic Legal Concepts [25] and the Core Legal Ontology based on the DOLCE foundational ontology [26]. Sophisticated logic-based languages for policy representation and reasoning on the Web include KAOs, Rei and Ponder [16], although the context was more the one of self-regulating intelligent systems rather than the expression of licence agreements. However, all these initiatives are still to be complemented by an equivalent effort in the legal domain, with the objective of providing the appropriate legal framework under which these technologies would operate, for example a *meta-rule of law* to give foundation to the adoption of metadata in supporting rights regulation on the Web of data [27, 28]. In this context, the term licence is often generalised with the notion of *policy* and in most of the following discussion we use them interchangeably. In what follows we give an overview of three popular approaches to licence expressions on the Web: MPEG, CC-REL, and ODRL.

Research on rights expression languages (REL) started in the early nineties, when the eXtensible Right Markup Language (XrML) was developed by Xerox, and later on included as reference language for the expression of rights in MPEG-21 [29]. Challenges for policy languages include (a) the capability to unambiguously define the terms and conditions of a policy (*policy expression*), (b) the assurance that all parties are mutually aware of the policy and its implications (*policy transparency*), (c) the potential to detect incompatibilities between policies, (d) methods to track exceptions and obligations [30].

Policies can be represented on the Web in a machine-readable format. The heterogeneity in licence descriptions raises the issue of modelling the licence descriptions themselves in a machine-readable way. Since the early investigation carried out on using RDF to police resource access [31], the landscape of licence models has witnessed the contributions of several actors in digital rights. MPEG's modelling for digital rights management is divided into the Rights Expression Language (REL) and the Rights Data Dictionary (RDD) [32]. A number of initiatives chose to apply a *semantic* approach for Digital Rights Management (DRM). Harmony project [33, 34] integrates copyright notions from the MPEG-21 RDD into a common ontology and OREL focuses on a formalising MPEG-21 RDD [14]. MPEG REL and RDD have been also integrated into a single web ontology framework for digital rights management [35]. MPEG-7 is a multimedia content description standard including a set of Description Schemes ("DS") and Descriptors ("D"), a language to express these schemes called the Description Definition Language (DDL), and an XML scheme for coding the description [36]. COMM is an attempt to develop a Web ontology guaranteeing that the intended semantics of MPEG-7 is fully captured and formalised, also using DOLCE as modelling basis [37].

The Creative Commons (CC) consortium publishes guidelines for describing permissions, jurisdictions and requirements on works in general. The Creative Commons Rights Expression Language was designed to express Creative Commons licences in a machine-readable format using RDF<sup>3</sup>. However, its expressivity is limited to the permissions, prohibitions and duties in the scope of CC licences. Specifically for data, the Open Data Institute has proposed the *ODRS* vocabulary<sup>4</sup>, which addresses licence compatibility and introduced the separation between data and content in the application of licences.

The Open Digital Rights Language (ODRL)<sup>5</sup> is a language to support the definition, exchange and validation of policies [38]. Recently, the W3C Permissions & Obligations Expression Working Group<sup>6</sup> followed up on ODRL to develop an official W3C standard for defining permissions and obligations, published as recommendation in 2018 [39].

<sup>3</sup>Creative Commons rights language, <https://creativecommons.org/ns>

<sup>4</sup>Open Data Rights Statement Vocabulary, <http://schema.theodi.org/odrs>

<sup>5</sup>ODRL Vocabulary & Expression, <https://www.w3.org/TR/2016/WD-vocab-odrl-20160721/>, Accessed: May, 2018.

<sup>6</sup>W3c Permissions & Obligations Working Group, [https://www.w3.org/2016/poe/wiki/Main\\_Page](https://www.w3.org/2016/poe/wiki/Main_Page), Accessed: May, 2018.

Recently, online repositories were developed to publish licences expressed in RDF, including *LicenseDB*<sup>7</sup>, which uses a mostly in-house vocabulary, and the RDFLicense Dataset of the Universidad Politécnica de Madrid<sup>8</sup> [40], and DALICC [5], which we use in our work.

In our work, we make the reasonable assumption that the policies used to express data licences are formulated with the expressivity of ODRL, in particular, the ODRL Web Ontology [38] that we outline in the next section.

### 3.3 Open Digital Rights Language (ODRL)

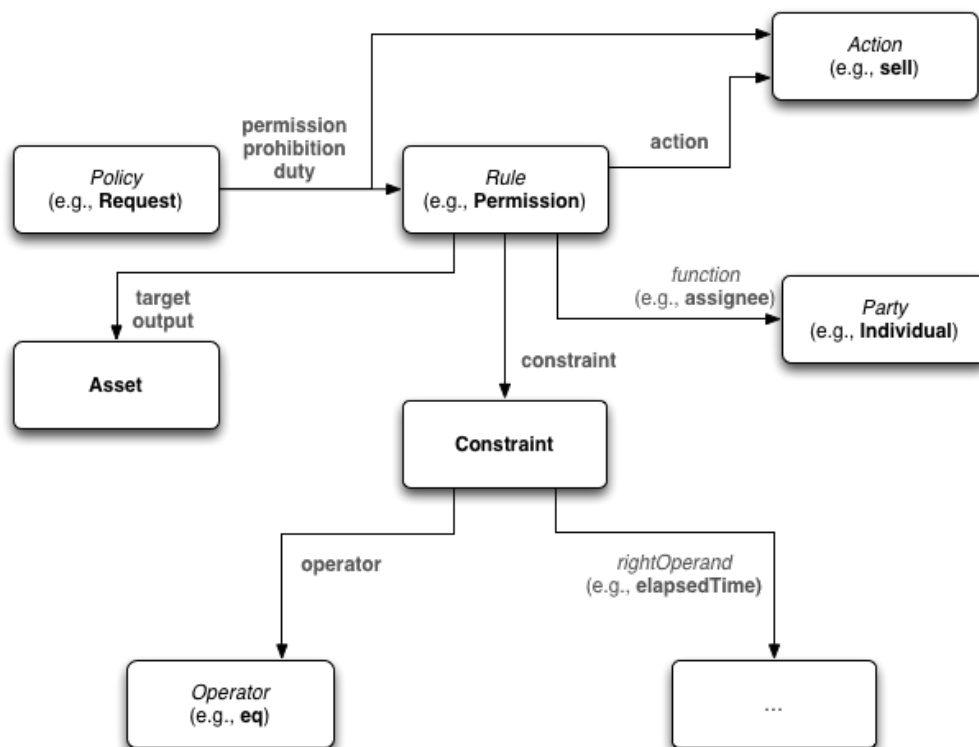


Figure 3.1: ODRL overview<sup>9</sup>.

The Open Digital Rights Language (ODRL) is a W3C standard curated and promoted by the related W3C ODRL Community Group<sup>10</sup>, which discusses use cases and requirements in support of enabling interoperability and transparent communication of policies associated with software, services, and data. The initial version was proposed for the XML language. Although ODRL is also available as an ontology, it only defines the semantics of policies in terms of natural language descriptions. An extension of the ODRL semantics has been proposed in [41] by considering

<sup>7</sup>LicenseDB, <http://licensedb.org>

<sup>8</sup>RDFLicense Dataset, <http://rdflicense.appspot.com/>, Accessed: May, 2018.

<sup>9</sup>Image taken from [38].

<sup>10</sup>W3C ODRL Community Group <https://www.w3.org/community/odrl/>, Accessed: May, 2018.

dependencies between actions, and discussing the impact of explicit and implicit dependencies on the evaluation of policy expressions.

The Open Digital Rights Language (ODRL), initially developed to support the management of digital rights, has evolved over the past years to a generic policy language (see Version 2.1 [38] and 2.2 [39]). Figure 3.1 shows the core entities of the underlying information model: *Asset*, *Policy*, *Action*, *Party*, *Rule*, and *Constraint*.

`odrl:Asset` is the top level entity type in ODRL, and represents any thing that can be subject to a policy (including a policy document itself). `odrl:Policy` is a generic class representing an entity meant for the expression of policies. A policy can be associated with an `odrl:Asset` with the `odrl:target` property. A policy can express an `odrl:Agreement`, `odrl:Offer`, and a `odrl:Request`. Moreover, a `odrl:Privacy` policy would refer to personal information contained in an `odrl:Asset`. Access to a given `Asset` can also be granted by the means of a `odrl:Ticket`, which allowance is represented as policy. In all other cases, a `odrl:Set` can be used to represent a composite policy. A `odrl:Action` can be object of a `odrl:permission`, `odrl:prohibition`, or `odrl:duty`. The language includes a vocabulary of instances of this class, including: `odrl:copy`, `odrl:user`, `odrl:derive`, `odrl:read`, `odrl:print`, `odrl:preview`, `odrl:distribute` and so forth. A policy can be associated with different parties (`odrl:Party`) depending on the role they play. For example, a party can be `odrl:assigner` or `odrl:assignee` of a policy, can be the one from which consent should be obtained - `odrl:consentingParty`, or the one to be informed - `odrl:informedParty`.

Assets can be then associated to licence documents by the means of the `odrl:target` property:

```
<http://purl.org/NET/rdflicense/OGL1.0> odrl:target <my-asset>
```

that is equivalent to using the Dublic Core term `dct:licence`: `dct:license` property:

```
<my-asset> dct:licence <http://purl.org/NET/rdflicense/OGL1.0>
```

While permissions, prohibitions and duties can be associated to single actions, more often they result from the combination of composite rules (`odrl:Permission`, `odrl:Prohibition`, or `odrl:Duty`) and constraints. ODRL can express the content of a data licence like the Open Government licence (OGL) described in the RDF Licences Database<sup>11</sup>, as showed in Listing 3.1.

### Listing 3.1: Open Government licence v1.0.

```
<http://purl.org/NET/rdflicense/OGL1.0> a odrl:Set;
dct:hasVersion "1.0" ;
rdfs:label "UK NonCommercial Government licence v1.0";
dct:language <http://www.lexvo.org/page/iso639-3/eng>;
cc:legalcode <http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/>;
odrl:permission [
  a odrl:Permission;
  odrl:action odrl:copy;
  odrl:action odrl:distribute;
  odrl:action odrl:derive ;
  odrl:duty [
    a odrl:Duty;
    odrl:action odrl:attribute;
    odrl:action odrl:attachPolicy ]] ;
odrl:prohibition [
  a odrl:Prohibition;
  odrl:action odrl:commercialize ] .
```

The **Data Licenses Clearance Center (DALICC)** [5] projects, that we reuse in our work, extends the set of ODRL

actions by covering new terms such as `dalicc:publish`, `dalicc:promote`, `dalicc:sublicense`, and others<sup>12</sup>. DALICC supports legal experts, businesses, and developers in the safe reuse of third-party digital assets such as datasets or software. Specifically, DALICC provides support for determining which asset can be shared with whom and under which conditions, thus lowering the burden of rights clearance. The system involves four components: a licence library of ready-made licences, a licence composer that reuses existing licence policies to create custom ones, a negotiator implementing tasks such as compatibility and conflict resolution, and a licence annotator to support the viewing of human-readable rights.

### 3.4 Reasoning with licences on the Web

Licences have an important role in regulating content policies on the Web. In this Section, we survey the problems and approaches for reasoning with licences.

A relevant area of application relates to software licences. Free and Open Source software (FOSS) gives the possibility to reuse third-party components in the development of new systems. Many different licences have appeared including various permissions and duties on the software use: GNU General Public licence (GPL), Apache License, MIT License, to name just a few. Generally speaking, the core objective of FOSS licences is to regulate distribution. As such, they are categorised in [42] as follows:

- (a) *permissive* - allowing software to be distributed under any licence (e.g. MIT, BSD, Apache licence 2.0);
- (b) *weak copyleft* - if the licenced software is modified then the resulting artefact must be distributed under the same licence. Also, if the derivation does not imply modification (but simple use) then any licence can be used (e.g. GNU Lesser General Public License, Mozilla Public License);
- (c) *strong copyleft* - any derivative work should use the same licence (e.g. GNU General Public License, Open Software License).

Software licences are typically assigned at component level. Normally, even small applications are composed of a large number of reused software components, and even determining all the licence involved in a given artefact is not a trivial task [43, 44]. Approaches are usually based on ad-hoc regular expressions on textual snippets extracted from source files or software archives [45, 46, 47].

Software reuse, repackaging and redistribution bring the problem of assessing whether the licences associated with the respective components are compatible. A given licence X is said to be one-way compatible with licence Y if the latter can be used in association with a software resulting as a combination of components released with licence X or Y [42]. As such, licence compatibility can be analysed by providing a graph-based model in which licences are nodes and arcs represent the one way, non-transitive, *compatible-with* relation (like in [48]). An alternative approach attempt to model the actual content of the licence with a specific meta-model [49]. Systems for supporting the construction, visualisation and comparison of arguments have been used in order to analyse the different software licences and compare their properties [50].

The Software Package Data Exchange (SPDX) specification developed in the context of the Linux Foundation<sup>13</sup> is promoted in order to standardise the way metadata about software components can be expressed and embedded in packaged releases, including the licences and copyrights associated. SPDX defines also a controlled set of identifiers for licences, listing more the 300 licences<sup>14</sup>. To date, licence selection in open source software repositories is proposed as a mere choice between a set of predefined licences, without any support (for example in SourceForge<sup>15</sup> and GitHub<sup>16</sup>). From our perspective, the limit of these approaches is that they are confined to the problems of source code and software reuse and redistribution. These approaches can be of limited use in understanding the actual content of the licences, as they are limited to the single dimension of redistribution of a derived artefact. Therefore, it is hard to assess to what extent they can be generalised.

<sup>12</sup>See <https://docs.dalicc.net/>.

<sup>12</sup>RDF Licences Database: <https://datahub.io/es/dataset/rdflicense>.

<sup>13</sup>SPDX: <https://spdx.org/>

<sup>14</sup>SPDX Licence list: <https://spdx.org/licenses/> - accessed 21st June 2017.

<sup>15</sup>SourceForge: <https://sourceforge.net/>

<sup>16</sup>GitHub: <https://github.com/>

In digital rights management (DRM), rule-based representation and reasoning over policies are required in order to enable secure data access and usage on the Web. Approaches can be grouped in the ones based on MPEG [32] and the ones relying on Semantic Web technologies [51, 52, 53]. MPEG-21 has been used as a reference model for performing licence compatibility assessment and recommendation, emerging from the scenario of aggregating open and private datasets in large smart city infrastructures [54]. A first order logic semantics for ODRL/XML has been proposed, and used to determine precisely when a permission is implied by a set of ODRL statements, showing that answering such questions is a decidable NP-hard problem [55]. Reasoning with ODRL has been also studied by defining its semantics according to *deontic logic*. In the latter, defeasible logic is used to reason with deontic statements, for example, to check compatibility of licences or to validate constraints attached to components on multi-agent systems [56].

Compatibility and composition are two fundamental problems with licenced content in the Web of data [57, 2, 58]. The problem of licences' compatibility has been extensively studied in the literature and tools that can perform such assessment do exist [59, 60]. Querying the Web of data implies integrating distributed data sources with a variety of policies attached to them. It is therefore problematic to assess the policies that should be associated with the query output, under the assumption that the query process is a way to combine the several data sources into a new one. This problem has been studied by applying two composition heuristics, AND-composition and OR-composition, and relying on ODRL and the rule-based reasoner SPINdle [59]. The deontic components specified by the source licences can be combined such to determine the acceptability of the operation (whether the policies of the combined data are conflicting or not) but also a new ad-hoc licence can be generated by combining the original policies [61]. Associating a licence to a dataset or service is a fundamental task when publishing on the Web. A formal representation of licences can be of use to support the users on deciding what possible constraints they want to guarantee concerning the use of their data. In [3], a formalisation of common data licences in ODRL is used to implement a tool to support the selection of a data licence, and to check the compatibility of one licence with another.

### 3.5 Large Language Models for Knowledge Extraction

The population and curation of such knowledge base is clearly a necessary step for licence recommendation systems. For example, the descriptions do not specify the types of assets a licence is eligible for (and we don't cover this aspect in the present paper). The enrichment of the possible terms to express policies will contribute to increase the precision and quality of the descriptions (see LiMO<sup>17</sup>, L4LOD<sup>18</sup> and ODRS<sup>19</sup>). Applying natural language processing techniques, like the ones proposed in [62], can facilitate the process of data acquisition. Recently, there has been increasing work on applying Large Language Models (LLM) to aid the extraction of structured information from textual content (e.g. [63, 64, 65]). Attention has been given on evaluating the suitability of LLMs in many end-user tasks as well as raising concerns on their limitations, for example, in generating plausible but wrong information (hallucination) and propagating societal biases derived from the text they have been trained from [66]. Knowledge graphs play a key role in bridging the gap between language models and structured data models [67], including attempts to mitigate known issues in content generated by LLMs such as hallucinations and biases [68]. Similarly, LLMs are at the centre of current effort in aiding knowledge graph population in various domains [69]. In our work, we use the power of LLMs to identify content from web resources that may include copyright and licence metadata, and develop a knowledge extraction pipeline that generates links between two knowledge graphs: one of resources and the other of licences represented computationally.

---

<sup>17</sup><http://data.opendataday.it/LiMo>

<sup>18</sup><http://ns.inria.fr/l4lod/v2/l4lodv2.htm>

<sup>19</sup><http://schema.theodi.org/odrs/>

## 4 Extracting licence information from web resources with a Large Language Model (LLM)

As discussed in Section 2.3, the musoW catalogue includes many resources for which a specific licence is not provided. The reasons for this may vary from the fact that such information does not exist (i.e. the resource manager does not provide a licence) to the possibility that annotators may overlook such information because it was hidden in secondary web pages. Certainly, the task of annotating license information is both laborious and prone to human errors. It's worth noting that dataset documentation may sometimes be incomplete or ambiguous, requiring direct communication with the dataset maintainers to clarify, as was the case for certain reuse experiences in Polifonia (the Choco knowledge graph [70]).

Large Language Models (LLM) such as OpenAI's ChatGPT, Meta's LLAMA and Google's Bard, have emerged recently providing impressive abilities in language generation, opening new opportunities for interacting with textual content, for example, for detecting and extracting structured information [71].

In this chapter, we report on experiments applying a Large Language Model (LLM) for extracting licence information from web resources to improve the coverage of the musoW licences metadata. Specifically, we pose the following questions:

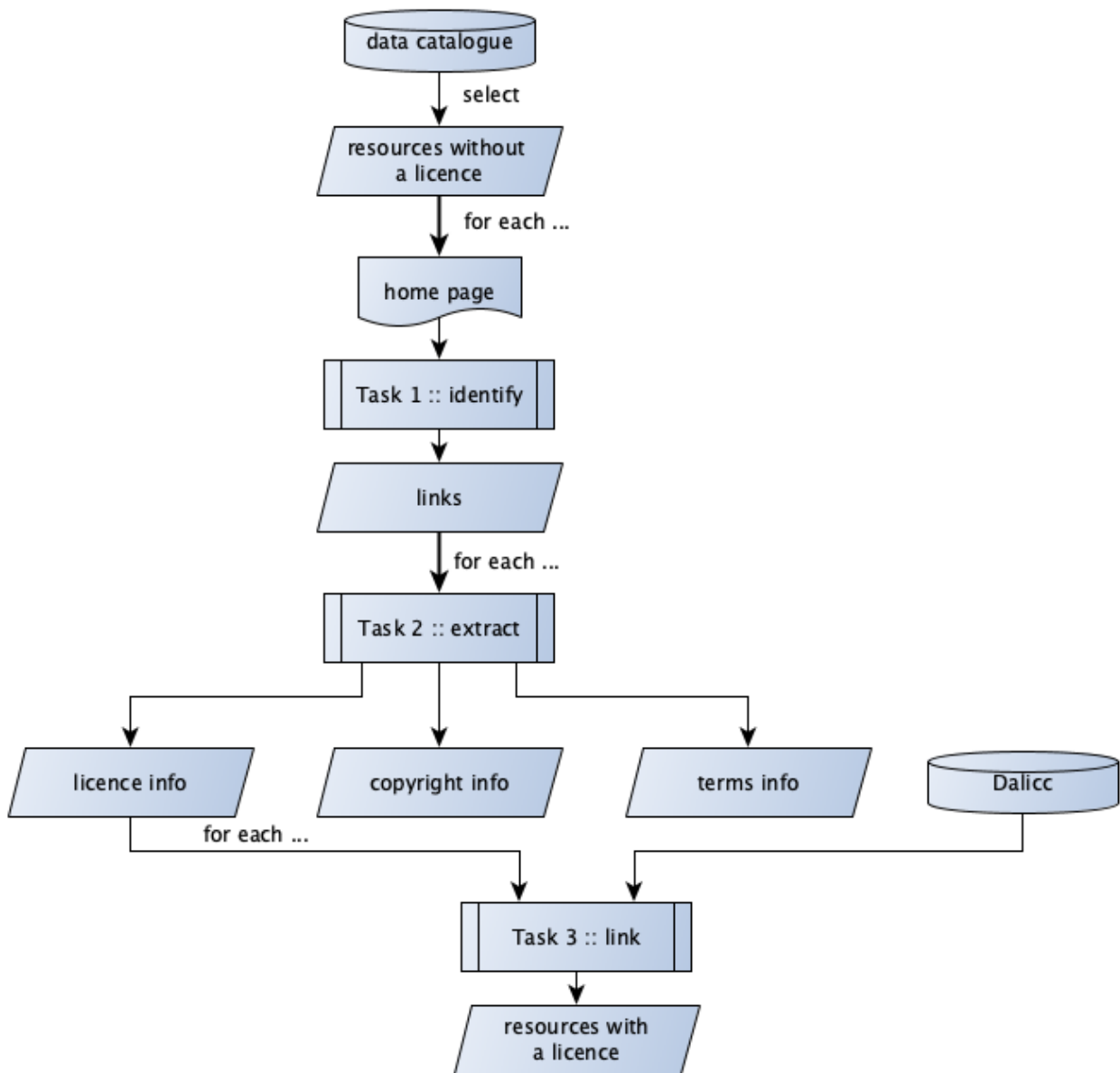
- RQ1 Can copyright and licence information be derived automatically from web pages?
- RQ2 How can copyright and licence information be derived automatically from web pages using Large Language Models (LLM)?
- RQ3 How accurate would an LLM detect the copyright and licence information (in other words, is it worth pursuing this line of enquiry)?
- RQ4 How much can we complete a curated catalogue of licence metadata with an automatic method based on LLMs?

### 4.1 Methodology

In this section, we describe our method for identifying, extracting, and linking licence information from web pages using an LLM. We tackle the problem by designing a methodology that engages with a large language model by asking to perform language understanding tasks. To avoid relying on the LLM embedded knowledge (which is known to be incomplete and often lead to unreliable information due to hallucinations), we design specific prompts that make use its language processing/predictive abilities but constraining them only to content that we provide.

The method, illustrated in Figure 4.1, is structured as follows:

- Data preparation We start from a list of resources published on the Web for which we want to know the associated licence. The assumption is that for each resource there is a web page which includes such information in one of the linked pages.
- Task 1 :: identify We design a prompt for an LLM asking to find no more than three links that may include copyright, privacy, or licencing information. The resource's home page is downloaded, and all HTML tags except anchor tags (links) are removed. This step is necessary for reducing the content size, making it less expensive to be analysed by the LLM. We ask the LLM to find such information in the content provided. The output of this process is a list of links potentially including copyright and licence information.
- Task 2 :: extract We design a prompt for an LLM asking to derive copyright, licence, and terms of use information from a piece of textual content. For each one of the resources and links collected, we download the HTML page and remove all tags. We then send the content to the LLM, which is asked to



**Figure 4.1:** Overview of the methodology

return a structured data object with three main fields: copyright, licence, and terms of use.

**Task 3 :: link** In this step, we focus on linking licencing information to a catalogue of well-known licences. We designed a prompt for an LLM asking to identify a licence from a piece of text, selecting it from a list provided.

**Evaluation** We evaluate each one of the previous steps under a number of dimensions, including 1. the ability of the LLM to provide an answer syntactically correct (following the requested specification); 2. the ability of the LLM to make an answer semantically correct (a meaningful answer);

Each task involving the LLM included a prompt engineering design phase which was essentially exploratory, starting from a prompt-as-hypothesis and resulting into a final prompt, after a short time of incremental trials with the LLM UI

dashboard. In what follows, we apply the methodology to the collection of resources in musoW which don't have a specified licence and describe our approach in detail.

## 4.2 Experiments

In this section, we report on the experiments conducted by applying the methodology outlined so far to the musoW catalogue of resources that do not have an explicit licence in the metadata. The experiments were executed using OpenAI ChatGPT API with model `gpt-3.5-turbo-16k`. The experiments are reproducible with the source code provided in this GitHub project: <https://github.com/polifonia-project/musow-licences-experiments-llm>.

### 4.2.1 Data preparation

The experiments are using two main resources: 1. the musoW catalogue of musical resources on the Web [6] 2. the DALICC catalogue of licences in RDF/ODRL [5].

We start by downloading the content from the musoW SPARQL endpoint<sup>1</sup>, specifically the resource identifier and name, the main home page of the resource, some categorical data and the licence metadata<sup>2</sup>. Next, we obtain the list of DALICC licences and generate a file summarising the licence description, legal text URL, and code used as a local name to identify the Linked Data entity<sup>3</sup>. The following listings show an excerpt of the file content, in YAML format:

```
- _position: 1
  code: NonCommercialGovernmentLicence
  link: http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/version/2/
  publisher: The National Archives
  title: Non-Commercial Government Licence
- _position: 2
  code: PythonLicense20
  link: http://opensource.org/licenses/Python-2.0
  publisher: Python Software Foundation
  title: Python License-2.0
- _position: 3
  code: Cc010Universal
  link: https://creativecommons.org/publicdomain/zero/1.0/legalcode
  publisher: Creative Commons
  title: CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
- _position: 4
  code: CC-BY_v4
  link: https://creativecommons.org/licenses/by/4.0/legalcode
  publisher: Creative Commons
  title: Creative Commons Attribution 4.0 International
```

**Listing 4.1:** Query run on the musow SPARQL endpoint to retrieve the list of resources and relevant metadata

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <https://schema.org/>
SELECT *
WHERE {
  ?resource
    schema:license ?licence .
  optional { ?resource schema:mainEntityOfPage ?webpage}
```

<sup>1</sup>musoW endpoint: <https://projects.dharc.unibo.it/musow/sparql>

<sup>2</sup>The data file can be inspected at <https://github.com/polifonia-project/musow-licences-experiments-llm/blob/main/Query-16.csv>.

<sup>3</sup>The YAML file can be found in the experiments project folder on GitHub: <https://github.com/polifonia-project/musow-licences-experiments-llm/blob/main/licences.yaml>



```
optional { ?resource rdfs:label ?label }
optional { ?resource schema:category ?category }
optional { ?resource schema:audience ?audience }
optional { ?resource schema:genre ?genre }
optional { ?resource a ?type }
optional { ?resource schema:additionalType ?type }
optional { ?resource schema:audience ?audience }
optional { ?licence rdfs:label ?licenceLabel }
}
```

#### 4.2.2 Task 1: finding links in web pages

The first task aims to automatically retrieve links pointing to web pages potentially including information about copyright, licence, and terms and conditions.

*Prompt engineering.* We start with the following prompt as an initial hypothesis:

SYSTEM: You are an expert in licencing and terms and conditions of resources on the Web.

USER: Find the link to the pages describing licences, privacy policies, or terms of use of the content in the following HTML source code. Please respond in a JSON format. HTML code: {{HTMLCODE}}

We perform tests with sample web pages from the musoW catalogue and change the prompt to include more details regarding the expected format and strengthen the reference to HTML knowledge. The resulting prompt is the following:

SYSTEM: You are expert in licencing and terms and conditions of resources on the Web. You also know how to find information on a web page by reading its HTML content.

USER: Find the link to the pages describing licences, privacy policies, or terms of use of the content in the following HTML source code. Please respond ONLY with a JSON format with a list of maximum 3 links, resolved according to this address: {url} HTML code: {html}

We iterate over the list of resources without explicit licence information (or marked with any of the categories that do not refer to a specific licence, as discussed in Section 2.3). The answers are saved locally and collected into a Table that we later analyse to evaluate the performance of the LLM under the two dimensions mentioned in our methodology, which we specify as follows:

Q1 Are there any links returned? (Yes/No)

Q2 Is the returned well-formed JSON? (Yes/No)

Q3\* Are any of those links relevant? We evaluate the answer on a Likert scale, from definitely not (1) to surely yes (5).

While the first two questions can be answered automatically, we rely on manual supervision to answer the third one (we indicate this with the asterisk). It needs to be duly noted that we did not manually check each one of the web pages but only observed the returned links and assessed whether any of them may potentially provide useful information. A sample of the results of this task can be seen in Table 4.1.

#### 4.2.3 Task 2: extract copyright, licence, and terms of use information from content

The output of the previous step is a set of links for each one of the resources derived from the content of the home web page. The second task aims to extract information from each one of those web pages. We used all links returned, independently from our manual relevance assessment (270 resources and 648 links in total).

**Table 4.1:** Example of results from Task 1. Relevance is established as follows: 1 - definitely irrelevant; 2 - probably irrelevant; 3 - cannot decide; 4 - probably relevant; 5 - definitely relevant.

web page	links found	rel.
<a href="http://www.discotheka.com/home/index?locale=en">http://www.discotheka.com/home/index?locale=en</a>		1
<a href="http://www.transforming-musicology.org/">http://www.transforming-musicology.org/</a>	<a href="https://transforming-musicology.org/">https://transforming-musicology.org/</a> <a href="https://www.themercurialmagpie.com/">https://www.themercurialmagpie.com/</a> <a href="https://plowns.com/">https://plowns.com/</a>	1
<a href="http://www.bruckner-online.at/">http://www.bruckner-online.at/</a>	<a href="http://www.bruckner-online.at/?page_id=604">http://www.bruckner-online.at/?page_id=604</a>	2
<a href="https://opac.rism.info/metaopac/start.do?View=rism&amp;Language=en">https://opac.rism.info/metaopac/start.do?View=rism&amp;Language=en</a>	<a href="https://www.bsb-muenchen.de/datenschutzerklaerung/">https://www.bsb-muenchen.de/datenschutzerklaerung/</a> <a href="https://www.bsb-muenchen.de/impressum/">https://www.bsb-muenchen.de/impressum/</a> <a href="http://www.rism.info/index.php?id=12&amp;L=0">http://www.rism.info/index.php?id=12&amp;L=0</a>	2
<a href="http://www.tudorpartbooks.ac.uk">http://www.tudorpartbooks.ac.uk</a>	<a href="http://www.ncl.ac.uk">http://www.ncl.ac.uk</a> <a href="http://www.tudorpartbooks.ac.uk/newsevents/">http://www.tudorpartbooks.ac.uk/newsevents/</a> <a href="http://www.tudorpartbooks.ac.uk/outputs/">http://www.tudorpartbooks.ac.uk/outputs/</a>	3
<a href="https://content.lib.washington.edu/wwwweb/">https://content.lib.washington.edu/wwwweb/</a>	<a href="https://content.lib.washington.edu/">https://content.lib.washington.edu/</a> <a href="https://digitalcollections.lib.washington.edu/cdm4/results.php">https://digitalcollections.lib.washington.edu/cdm4/results.php</a> <a href="http://www.womenwhorockcommunity.org">http://www.womenwhorockcommunity.org</a>	3
<a href="https://www.metal-archives.com/">https://www.metal-archives.com/</a>	<a href="https://www.metal-archives.com/content/help">https://www.metal-archives.com/content/help</a>	4
<a href="http://www.rolldabeats.com/">http://www.rolldabeats.com/</a>	<a href="https://www.rolldabeats.com/forum/index.php?/contact/">https://www.rolldabeats.com/forum/index.php?/contact/</a> <a href="https://www.rolldabeats.com/forum/index.php?/cookies/">https://www.rolldabeats.com/forum/index.php?/cookies/</a>	4
<a href="http://www.musicatradicional.eu">http://www.musicatradicional.eu</a>	<a href="http://www.musicatradicional.eu/contact">http://www.musicatradicional.eu/contact</a>	4
<a href="http://www.sjsu.edu/beethoven/">http://www.sjsu.edu/beethoven/</a>	<a href="http://www.sjsu.edu/privacy/index.html">http://www.sjsu.edu/privacy/index.html</a> <a href="http://www.sjsu.edu/titleix/">http://www.sjsu.edu/titleix/</a> <a href="http://www.sjsu.edu/siteindex/">http://www.sjsu.edu/siteindex/</a>	5
<a href="http://www.ccarh.org/">http://www.ccarh.org/</a>	<a href="http://www.ccarh.org/software/dmuseviewer">http://www.ccarh.org/software/dmuseviewer</a> <a href="http://www.ccarh.org/software/musedata/">http://www.ccarh.org/software/musedata/</a> <a href="http://www.ccarh.org/privacy-policy">http://www.ccarh.org/privacy-policy</a>	5
<a href="http://www.lib.uchicago.edu/e/su/cja/sheetmusic.html">http://www.lib.uchicago.edu/e/su/cja/sheetmusic.html</a>	<a href="https://www.lib.uchicago.edu/about/thelibrary/policies/">https://www.lib.uchicago.edu/about/thelibrary/policies/</a> <a href="https://www.lib.uchicago.edu/research/help/ask-librarian/ask-contact/">https://www.lib.uchicago.edu/research/help/ask-librarian/ask-contact/</a> <a href="https://www.lib.uchicago.edu/scrc/finding-aids/?topic=Jazz&amp;view=topics">https://www.lib.uchicago.edu/scrc/finding-aids/?topic=Jazz&amp;view=topics</a>	5

For this task, we want the information to be structured under three dimensions: copyright statement – who owns the intellectual property of the resource; licence – what is the licence associated with it (if any); and terms of use – to include any other information regarding the use of the resource.

*Prompt engineering.* We start with the following prompt as an initial hypothesis:

SYSTEM: You are expert in licencing and terms and conditions of resources on the Web. You also know how to find information on a web page by reading its HTML content.

USER: Please list the licences and copyright owners named in the following HTML code. Format the answer in JSON with two fields, 'copyright' and 'licences'.  
 {{HTMLCODEE}}

We perform tests with a sample of content from the web pages of the previous step and refine the prompt until we obtain sufficiently consistent results. The resulting prompt is the following:

SYSTEM: You are expert in licencing and terms and conditions of resources

on the Web. You also know how to find information on a web page by reading its HTML content and express it in JSON format.

USER: Please list the licences, copyright owners, and terms and conditions mentioned in the following text. Respond only with a JSON object with 3 fields, 'copyright', 'licences', and 'terms and conditions'. The text is: {text}

We save the responses locally and gather them in a tabular format. Subsequently, we analyze this data to assess the effectiveness of the LLM based on the two dimensions stated in our methodology. These dimensions are elaborated as follows:

- Q4 Is the returned well-formed JSON?
- Q5,Q8 Did the LLM find any copyright information?
- Q6,Q9 Did the LLM find any licence information?
- Q7,Q10 Did the LLM find any terms and condition information?

We pose the last three questions above two times, the first considering each of the links (web pages) and associated requests to the LLM and the second aggregating all responses related to each resource and quantifying whether any provided links was useful to gather the information. While the above questions can be answered automatically, we add a qualitative, human-based assessment of the quality of the results, answering the following additional questions on a restricted sample of 100 items:

- Q11\* Is the copyright information correct?
- Q12\* Is the licence information correct?

Tables 4.3 and ?? show example annotations for questions Q11 and Q12 respectively. Crucially, we observed that for all 100 evaluated responses to Q12, the LLM never returned a wrong answer, while having some variability in the form (for example, in some cases it did not find a licence but it still returned some content). We leave the assessment of the information related to the terms and conditions to future work.

**Table 4.2:** Example annotations for evaluating Q11: 0 - No copyright; 1 - Somehow correct; 2 - Surely correct.

Web page	Copyright	Ann.
<a href="https://github.com/midi-ld/documentation/issues">https://github.com/midi-ld/documentation/issues</a>	2023 GitHub, Inc.	0
<a href="https://github.com/midi-ld/">https://github.com/midi-ld/</a>	GitHub, Inc.	0
<a href="http://drm.ccarh.org">http://drm.ccarh.org</a>	Unknown	0
<a href="https://www.muziekweb.nl/Informatie/Privacybeleid">https://www.muziekweb.nl/Informatie/ Privacybeleid</a>	Stichting Nederlands instituut voor Beeld & Geluid, 1995 - 2023	0
<a href="http://popmusic.mtsu.edu/ManuscriptMusic/guidelines.aspx">http://popmusic.mtsu.edu/ManuscriptMusic/guidelines.aspx</a>	Copyright © 2023 – All Rights Reserved	1
<a href="https://library.duke.edu/using/policies/dmca-takedown">https://library.duke.edu/using/policies/dmca-takedown</a>	Duke University	1
<a href="https://archive.list.co.uk/the-list/2020/index.html">https://archive.list.co.uk/the-list/2020/index.html</a>	The List Archive	1
<a href="https://www.uni-regensburg.de/impressum/index.html">https://www.uni-regensburg.de/impressum/index.html</a>	Universität Regensburg	1
<a href="https://www.youtube.com/t/privacy">https://www.youtube.com/t/privacy</a>	Google LLC	2
<a href="https://aaamc.indiana.edu/whats-going-on/aaamc-speaks/index.php">https://aaamc.indiana.edu/whats-going-on/aaamc-speaks/index.php</a>	The Trustees of Indiana University	2
<a href="http://cantus.uwaterloo.ca">http://cantus.uwaterloo.ca</a>	Cantus Database © 2012-2022	2
<a href="http://pemdatabase.eu/">http://pemdatabase.eu/</a>	Portuguese Early Music Database (PEM), 2010-2023	2
<a href="http://cantus.sk">http://cantus.sk</a>	(c) Slovak Early Music Database 2012-2023	2

**Table 4.3:** Example annotations for evaluating Q12. The **F?** columns evaluates whether the LLM found a licence in the content: 0 - Not found; 1 - Found but incorrect; 2 - Found and correct. The **Inc?** column evaluates if the response was not correct.

Web page	Licence	F?	Inc?
http://popmusic.mtsu.edu/ManuscriptMusic/guidelines.aspx	[]	0	0
http://www.mcgill.ca/music/		0	0
https://www.youtube.com/t/terms	worldwide, non-exclusive, royalty-free, transferable, sub-licensable licence to use that Content	0	0
https://web.library.yale.edu/data-use	[]	0	0
https://github.com/midi-ld/	['MIT']	2	0
http://cantus.uwaterloo.ca	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License	2	0
http://pemdatabase.eu/	['Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)']	2	0
https://libraries.mit.edu/permissions	['CC BY-NC']	2	0

#### 4.2.4 Task 3: link licence descriptions to the licences database

The expected output of the previous step is a structured JSON object with three fields: copyright, licence, and terms of use. In this task, we focus on the content returned for the field 'licence' and aim at automatically link such licence descriptions with the equivalent authoritative entry derived from the Dalicc catalogue of licences expressed in RDF/ODRL (that we introduced in Section 3.3). The initial prompt hypothesis is the following:

```
SYSTEM: You are expert in licencing and terms and conditions of resources on
the Web. You also know how to find information on a web page by reading
its HTML content. You are also proficient in reading YAML files.
USER: Given the following list of licences, can you tell me to which
licence the following description refers to {LICENCEEXPR} {YAML}
```

We refined the prompt by testing a sample of licence descriptions identified in the previous step to improve the results. Specifically, we moved the list of licences to the SYSTEM input and asked to return 'NONE' if the description would not refer to any specific licence in the list.

```
SYSTEM: You are expert in licencing and terms and conditions of resources on
the Web and know the following list of licences:
{listOfLicences}
USER: Can you tell me to which licences the following licence description
refers to? The description is {description} -- Please respond by only
reporting the selected licences from the list or 'NONE' if none is found.
```

We manually evaluate each one of the responses and annotate them as follows:

- 1 The licence described is in the list, but the LLM didn't find it (or it hallucinated in some way)
- 0 The licence described is not in the list and the LLM correctly did not find it

- 1 The licence description found is correct and in the list but the LLM did not link it properly (for example, it did not respond with the licence code)
- 2 The licence was found in the list and linked properly (the correct licence code was returned)

Table 4.4 shows a sample of annotated responses, limited to the content of Task 2 returning the field 'licence'.

**Table 4.4:** Examples from the results of task 3.

'licence' description from Task 2	answer of the LLM to Task 3	ann.
[["{'licence': 'Library of Congress'}, {'licence': 'ReadSpeaker'}, {'licence': 'Facebook'}, {'licence': 'Twitter'}, {'licence': 'Bachtrack'}, {'licence': 'WETA Classical Conversations'}, {'licence': 'National Public Radio'}]]"]	The licences corresponding to the descriptions are:- Library of Congress: [CC0 1.0 Universal]- ReadSpeaker: [CC-BY-ND_v4]- Facebook: [CC-BY_v4]- Twitter: [CC-BY_v4]- Bachtrack: [CC-BY_NC-ND_v4]- WETA Classical Conversations: [CC-BY_NC-ND_v4]- National Public Radio: [CC-BY_v4]	-1
[["{'ELVIS Database Source': 'https://github.com/ELVIS-Project/elvis-database/releases'}]]"]	NONE	0
[["{'MIT'}]]"]	The description 'MIT' corresponds to the [ExpatLicense] MIT License (MIT) <a href="https://opensource.org/licenses/MIT">https://opensource.org/licenses/MIT</a>	1
[["{'title': 'CC BY-NC 4.0', 'website': 'https://creativecommons.org/licenses/by-nc/4.0/'}]]"]	[CC-BY-NC_v4] Creative Commons Attribution-NonCommercial 4.0 International (Creative Commons) <a href="https://creativecommons.org/licenses/by-nc/4.0/legalcode">https://creativecommons.org/licenses/by-nc/4.0/legalcode</a>	2
[["{'name': 'Creative Commons Attribution License', 'url': 'http://creativecommons.org/licenses/by/1.0/'}]]"]	CC-BY_v4	2

We quantitatively evaluate this task as follows:

- Q13\* How many correct decisions are made? (all except -1)  
 Q14\* How many licences are correctly not found? (0)  
 Q15\* How many licences are correctly found? (1 and 2)  
 Q16\* How many licences are linked to the list? (2)

Q13 includes all answers, positive and negative, while Q14 summarise the licences that were missing from the sources and therefore not linked. Q15 counts the licences that were found and correct, even if the linking task didn't work syntactically, while Q16 only measures the licences that were correct and properly linked to the list provided.

### 4.3 Results and discussion

In this section we present the results of our experiments and discuss them in the light of the initial assumptions and hypotheses. A summary of the results is reported in Table 4.5. The results of each one of the steps, the related manual annotations, and the computed statistics can be reviewed at this address: <https://docs.google.com/spreadsheets/d/1wl-5YKcLVY9wDwSauPWz9NlYel7Ga1Da5WJXtOOp18/edit?usp=sharing>.

We can first look at the results of each one of the tasks, in order to gather evidence that would allow us to answer the main questions.

**Task 1** The first task is related to finding links in web pages that may include copyright or licence information. The task was executed 313 times, one for each resource home page. The vast majority of results were provided with a

correct JSON syntax (this includes responses with no links). The LLM was capable of finding links in 86% of the cases, and most of the link sets are deemed to be potentially relevant (51% were surely relevant and 21% were deemed potentially relevant by our manual assessment).

**Task 2** The second task aims at extracting textual content from the web pages mentioning copyright, licence, or terms of use information. The task was executed 648 times, one for each web page collected in the previous step. Those links covered 86% of the collection (270). A good amount of results were provided with a correct JSON syntax – 75% (this includes responses with no information). Copyright information was found in 66% of the cases (82% of the resources, 221/270), while licence information had a much lower result: 26%, corresponding to less than half of the resources for which at least one web page was returned in the previous step (43%). Terms of use are also found with a similar success rate, however, we don't delve into those now and leave an assessment of the quality of this additional information to future work.

At the end of this second step, out of 313 initial resources, we obtain copyright information for 221 of them and licence information for 115 of them, approximately 70% and 36% respectively. The reasons vary from errors propagated from the previous step to the information not existing at all in the web pages. Crucially, we validate the quality of the results with a manual supervision of a sample of 100 resources, for which we find that 65% include correct copyright information and 100% include correct licence information (or did not find any when none was there). This information was checked by manually opening each one of the web page and verifying its content. Crucially, we can be confident that the LLM did not hallucinate when requested to derive licence information from a web page, therefore, that the returned content, when valid, is also true.

**Task 3** The last task is devoted to automatically link the licence information to the list of licences of the Dalicc catalogue. The results of this operation, performed on the 115 resources that included any form of licence information (including cases where such information was empty, missing, or non-referring to a specific licence). We evaluate the entire result set manually according to a Likert scale of 5, reflected in questions 13-16 (see Table 4.5). The prompt to the LLM was to identify licences from the list provided, starting from a text that supposedly mentions any of them. We can observe how the system made a correct decision (whether there was a licence from the list or not) in 90% of the cases. However, in more than half of the cases, there was no licence information – 57%. However, the system managed to correctly identify a licence from the Dalicc catalogue for 38 resources (33% of the cases) and in 25% of the cases it was able to report the correct licence code from the list (76% of the ones correctly found).

With this approach, we managed to retrieve and link 38 licence information in an automatic (or semi-automatic) way, covering 12% of the resources which originally did not have a licence specified.

We conclude this section by discussing the original research questions.

**[RQ1] Can copyright and licence information be derived automatically from web pages?** We can conclude that it is possible to derive such information from web pages, and automatic method involving LLM can help in processing large amount of web pages and to gather relevant information with little human supervision. Crucially, we gathered evidence that there is little risk to generate plausible but wrong information in the case of licencing, thus making us confident that it is possible to apply LLM for extracting licencing information from the content of web resources (see Table ??). This is not true for copyright, as shown by our evaluation of Q11 (reported in Table 4.3).

**[RQ2] How can copyright and licence information be derived automatically from web pages using Large Language Models (LLM)?** Our methodology, which was validated by our experiments, is an initial answer to this question. However, we performed our experiments with one specific LLM (ChatGPT) and we acknowledge the fact that a larger study would be needed in order to establish what kind of prompts would be most successful generally, for example considering portability across different LLMs, in achieving this task. However, our experiments are promising and open directions about how to improve the overall workflow both in terms of accuracy and coverage.

**[RQ3] How accurate would an LLM detect the copyright and licence information (in other words, is it worth pursuing this line of enquiry)?** By looking into the results, we can observe how most of the decrease of coverage during the pipeline was due either to difficulties in producing machine-readable content or in actually recognise that the information is not there (for example, this can be seen by comparing the results of Q9 with Q14). Increasing correct responses in the case of true negatives seems to be a challenge (sometimes the LLM returns some content that is not including relevant information in a task but then this becomes ineffective in further tasks, for example when the LLM returns a piece of text that does not describe a licence in Task 2 and the same text is correctly not linked to any licence in Task 3). Instead we can observe how the LLM was particularly accurate in deciding, for example, whether a certain piece of text was including a licence from a given list (Q13). These results are particularly encouraging and we can definitely see this as a promising research direction.

**[RQ4] How much can we complete a curated catalogue of licence metadata with an automatic method based on LLMs?** This final answer pertains to our case studies. We managed to find new licence information for 38 resources (12% of the set of resources without licence annotations). We cannot confidently state that those are all the existing missing ones but from the analysis of the results of intermediate steps in our pipelines, we are confident that most of the web pages scrutinised did not include licence information (see results about Q12 and Q13). This is also coherent with the original statistics in musoW, where most of the resources did not present licence information. However, our method allowed us to get more of them, inspiring is in considering opportunities for adopting LLM as an aid for curating digital libraries metadata.

**Table 4.5:** Summary of the results

<b>Qn</b>	<b>Task</b>	<b>Question</b>	<b>True</b>	<b>Maybe</b>	<b>False</b>	<b>Total</b>	<b>% True</b>
1	T1	Are there any links returned?	270	0	43	313	0.86
2	T1	Is the returned well-formed JSON?	293	0	20	313	0.94
3	T1	Are any of those links relevant? (*)	160	67	86	313	0.51
4	T2	Is the returned well-formed JSON?	485	0	163	648	0.75
5	T2	Did the LLM find any copyright information?	428	0	220	648	0.66
6	T2	Did the LLM find any licence information?	169	0	479	648	0.26
7	T2	Did the LLM find any terms information?	235	0	413	648	0.36
8	T2	Did the LLM find any copyright information? (R)	221	0	49	270	0.82
9	T2	Did the LLM find any licence information? (R)	115	0	155	270	0.43
10	T2	Did the LLM find any terms information? (R)	133	0	137	270	0.49
11	T2	Is the copyright information correct? (*)	47	0	16	63	0.75
12	T2	Is the licence information correct? (*)	26		0	26	1.00
13	T3	How many correct decisions are made? (all except -1)	104	0	11	115	0.90
14	T3	How many licences are correctly not found? (0)	66	0	49	115	0.57
15	T3	How many licences are correctly found? (1 and 2)	38	0	77	115	0.33
16	T3	How many licences are linked to the list? (2)	29	0	86	115	0.25
17	-	How many licences are automatically linked (of the ones correctly found)?	29	0	9	38	0.76
18	-	How many licences are linked (of the total resources)?	38			313	0.12



## 5 Knowledge Graph Construction

In this section we describe the production of the Polifonia Licences Knowledge Graph, which integrates the musoW knowledge graph, complemented with the results of the LLM-powered knowledge extraction pipeline described in Chapter 4, with the licences of the Dalicc catalogue expressed in RDF/ODRL.

The integration process we describe below is lead by SPARQL Anything. Before proceeding with detailing the pipeline, we briefly introduce the tool. Recent studies on constructing knowledge graphs suggest using a concept called *façade* [72], borrowed from object-oriented software engineering. This concept is applied to SPARQL Anything, which enables querying heterogeneous resources *as-if* they were in RDF using plain SPARQL 1.1 through the overloading of the SERVICE clause. SPARQL Anything supports a wide range of file formats, from commonly used ones like CSV, JSON, XML, and Spreadsheets to others that are not supported by other solutions such as Markdown, YAML, DOCx, and Bibtex. It also offers features such as flexible querying of Web APIs, parametrized queries, and the ability to chain multiple transformations into complex pipelines. The system is developed by a community of contributors as open source software, with core contributions by the EU-funded projects SPICE and Polifonia. The project is managed on GitHub at this address: <https://github.com/SPARQL-Anything>, and can be cited via its related entry in Zenodo [73], following good practices of Open Science and FAIR data management policies. The official documentation is published via Readthedocs.io<sup>1</sup>.

The data integration pipeline described in this section is publicly available for reuse at <http://github.com/polifonia-project/licences>. In what follows, we use the SPARQL Anything command line interface, referring to it as the tool `fx`. The process is divided in the following steps:

1. Harvesting RDF/ODRL descriptions from the DALICC licences catalogue
2. Downloading MusoW
3. Linking musoW licence metadata to DALICC identifiers
4. Integrating licences generated by the experiments (see Chapter 4)

**Harvesting RDF/ODRL descriptions from the DALICC licences catalogue** We reuse a catalogue of machine readable licences from the [Dalicc project](<https://www.dalicc.net/>). We use SPARQL Anything to download the licence representation from each one of the items in the dalicc catalogue (see Listing 5.1, executed as follows:

```
fx -q queries/harvest-dalicc.sparql -f TTL -o knowledgegraph/dalicc.ttl
```

**Listing 5.1:** The SPARQL Anything query to harvest licence descriptions from the dalicc catalogue.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX odrl: <http://www.w3.org/ns/odrl/2/>
PREFIX fx: <http://sparql.xyz/facade-x/ns/>
PREFIX xyz: <http://sparql.xyz/facade-x/data/>

CONSTRUCT {
    ?s ?p ?o
} WHERE {
    service <x-sparql-anything:> {
        service <x-sparql-anything:> {
            fx:properties fx:location "https://api.dalicc.net/licenselibrary/list" ;
            fx:http.query.limit 10000 ;
            fx:media-type "application/json"
```

<sup>1</sup><https://sparql-anything.readthedocs.io/en/latest/>

```

        .
    [] xyz:id/xyz:value ?uri
    }
    BIND (CONCAT(?uri, "?format=ttl&download=true") as ?location) .
    fx:properties fx:location ?location ;
        fx:http.header.accept "text/turtle" ;
        fx:http.query.format "ttl" ;
        fx:http.query.download "true" ;
        fx:media-type "text/turtle"
    .
    ?s ?p ?o
}
}

```

**Downloading musoW** Obtaining a copy of the musoW catalogue is easy via the SPARQL endpoint:

```

CONSTRUCT {
  ?x ?y ?z
} WHERE {
  SERVICE <https://projects.dharc.unibo.it/musow/sparql> {
    ?x ?y ?z
  }
}

```

The query can be executed with our tool in the same way as any SPARQL query:

```
fx -q queries/download-musow.sparql -f TTL -o knowledgegraph/musow.ttl
```

The output is a local snapshot of the musoW KG.

**Linking musoW licence metadata to DALICC identifiers** A list of licence annotations for musoW was presented in Section 2.3. We aligned the licence descriptions to the equivalent Dalicc identifiers, when possible, as follows:

```

@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix dalicc: <http://dalicc.net/licenseslibrary/>.
@prefix musow: <https://w3id.org/musow/vocab/>.

musow:cc-by owl:sameAs dalicc:CC-BY_v4 .
#musow:open-access a
musow:cc-by-3.0 owl:sameAs dalicc:CC-BY_v4 .
musow:cc-by-4.0 owl:sameAs dalicc:CC-BY_v4 .
musow:cc-by-nc owl:sameAs dalicc:CC-BY-NC_v4 .
musow:cc-by-nc-nd owl:sameAs dalicc:CC-BY-NC-ND_v4 .
musow:cc-by-nc-sa owl:sameAs dalicc:CC-BY-NC-SA_v4 .
musow:cc-by-sa owl:sameAs dalicc:CC-BY-SA_v4 .
musow:cc0 owl:sameAs dalicc:Cc010Universal .
#musow:copyright
musow:gnu-fdl owl:sameAs dalicc:GnuFreeDocumentationLicense .
musow:gnu-gpl owl:sameAs dalicc:GNU_GPL_v3 .
musow:mit-license owl:sameAs dalicc:MIT .
#musow:not-specified
#musow:privative
#unknown-licence

```

Then, we materialised triples according to the ODRL vocabulary with the following query:

```

CONSTRUCT {

```

```
?resource a odrl:Asset
.
    ?daliccLicence odrl:asset ?resource .
} WHERE {
  graph ?g { ?resource schema:license ?li1 }
  graph ?k { ?li2 owl:sameAs ?daliccLicence }
  filter (?li1 = ?li2)
}
```

executed as follows:

```
fx -q queries/musow-licences.sparql -f TTL -o knowledgegraph/musow-licences.ttl
```

**Integrating licences generated by the experiments** However, musoW licence annotations are complemented with additional metadata from [experimenting with extracting and linking licence information from web resources with the help of LLMs](as discussed in Chapter 4).

In summary, the Polifonia Licences Knowledge Graph is composed of the following files (and related graphs):

- dalicc.ttl
- musow.ttl
- musow-alignments.ttl
- musow-licences.ttl
- musow-licences-llm.ttl

## 6 Querying the Polifonia Licences Knowledge Graph

In this chapter we query the knowledge graph to answer information needs discussed in Chapter 2. The knowledge graph includes detailed informations about permissions, prohibitions, and duties associated with 130 resources from the musoW catalogue, including the data used and produced by the Polifonia project.

Table 6.1 summarises the licences included in the KG.

**Table 6.1:** Statistic of licences in the knowledge graph

id	name	resources
CC-BY_v4	Creative Commons Attribution 4.0 International	55
Cc010Universal	CC0 1.0 Universal (CC0 1.0) Public Domain Dedication	28
CC-BY-NC-SA_v4	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International	16
CC-BY-NC_v4	Creative Commons Attribution-NonCommercial 4.0 International	11
GNU_GPL_v3	GNU General Public License 3	8
CC-BY-SA_v4	Creative Commons Attribution-ShareAlike 4.0 International	7
MIT	The MIT License	2
Apache-2.0	"Apache License Version 2.0"	1
CC-BY-NC-ND_v4	Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International	1
GnuFreeDocumentationLicense	GNU Free Documentation License version 1.3	1

Thanks to a computational representation of terms of use linked to the resources of the musoW catalogue, we are now able to answer key questions in relation to permissions, prohibitions, and duties relevant to musical resources' reuse, that we discuss. Table 6.2a shows a summary of actions and the number of resources associated with them. In what follows, we discuss how each one of the competency questions introduced in Chapter 2 can be now answered by the knowledge graph.

**[CQ1] Do you want to reuse an external dataset?** In the case pilot developers need to reuse a dataset, they can query The Polifonia Licence Knowledge Graph for resources that allow derived content. Listing 6.1a shows a SPARQL query for this use case. The query counts the number of resources for which it is possible to derive new content (the results are shown in Table 6.2b).

**[CQ2] Does the licence covering the dataset permit to share?** Similarly to the previous CQ, we can query for resources whose licence include the `odrl:action odrl:distribute` (see also Table 6.2a).

**[CQ3] If yes, what are the licence requirements for sharing the data?** Licence requirements for sharing can be queried by asking for all duties associated with a permission to share. We can query the knowledge graph for the restrictions associated with resources with the SPARQL code illustrated in Figure 6.1b.

The result of this query is reported in Table 6.3 (limited to licences associated with more than 2 assets).

**Table 6.2: Query results**

(a) Summary of actions and the number of resources associated with them

type	operation	resources
Can	ChangeLicense	29
Can	CommercialUse	101
Can	DerivativeWorks	127
Can	ModifiedWorks	128
Can	chargeDistributionFee	96
Can	derive	127
Can	display	128
Can	distribute	128
Can	modify	128
Can	present	128
Can	reproduce	128
Cannot	ChangeLicense	97
Cannot	CommercialUse	28
Cannot	chargeDistributionFee	1
Cannot	derive	1
Cannot	promote	126
Must	Attribution	100
Must	Notice	100
Must	SourceCode	9
Must	compliantLicense	1
Must	modificationNotice	100
Must	source	1

(b) Results of query related to CQ1.

assets	licence
62	CC-BY_v4
36	Cc010Universal
25	CC-BY-NC-SA_v4
16	CC-BY-NC_v4
13	CC-BY-SA_v4
13	GNU_GPL_v3
3	MIT
2	CC-BY-NC-ND_v4
2	GnuFreeDocumentationLicense
1	Apache-2.0
1	LGPLv3

**[CQ4] Do you want to modify the existing dataset? and [CQ5] Does the licence covering the dataset permit to modify the original data?** The Polifonia Licence Knowledge Graph can answer to these questions by leveraging the `odrl:action modify`, as illustrated by Table 6.2a.

**[CQ6] If yes, what are the licence requirements for modifying the data?** Depending on how we interpret this competency question, we may refer to the possibility of producing derived works, modifying the original data, or asking for the restrictions in relation to derived content. All three of these interpretations are covered by the ODRL descriptions of the knowledge graph. Derivative or modified data are allowed in 127 and 128 of the cases, respectively. See Table 6.3.

**[CQ7] Does the licence impose a ShareAlike obligation?** Table 6.2a shows how 97 resources have a licence that does not allow to distribute derived content with a different licence (Cannot ChangeLicense).

**[CQ8] Does the licence impose any commercial restrictions?** Table 6.2a shows how 101 resources have a licence that explicitly permits commercial use (Can CommercialUse), while 28 have a licence that explicitly forbids it (Cannot CommercialUse).

**Figure 6.1: Queries.**

**(a) Query to answer CQ1.**

```

PREFIX odrl: <http://www.w3.org/ns/odrl
/2/>
PREFIX afn: <http://jena.apache.org/ARQ/
function#>

SELECT (COUNT(?asset) AS ?assets)
      (afn:localname(?licence_) as ?licence
)
WHERE {
    graph ?gMusow {
        ?licence_ odrl:asset ?asset
    }
    graph ?gLicences {
        ?licence_ odrl:permission [odrl:
            action odrl:distribute]
    }
}
group by ?assets ?licence_
order by DESC(?assets) ?licence_

```

**(b) SPARQL query to retrieve the restrictions associated with resources.**

```

PREFIX odrl: <http://www.w3.org/ns/odrl
/2/>
PREFIX afn: <http://jena.apache.org/ARQ/
function#>

SELECT (COUNT(?asset) AS ?assets)
      (afn:localname(?licence_) as ?licence
)
      (afn:localname(?can_) as ?can)
      (afn:localname(?must_) as ?must)
WHERE {
    graph ?gMusow {
        ?licence_ odrl:asset ?asset
    }
    graph ?gLicences {
        ?licence_ odrl:permission ?perm .
        ?perm odrl:action ?can_ ;
            odrl:duty/odrl:action ?must_
    }
}
group by ?assets ?licence_ ?can_ ?must_
order by DESC(?assets) ?licence_ ?can_ ?
must_

```

**Table 6.3:** Results of the query answering CQ5 and CQ6.

Assets	Licence	Can	Must
assets	licence	can	must
62	CC-BY_v4	derive	Attribution
62	CC-BY_v4	derive	Notice
62	CC-BY_v4	derive	modificationNotice
62	CC-BY_v4	distribute	Attribution
62	CC-BY_v4	distribute	Notice
62	CC-BY_v4	modify	Attribution
62	CC-BY_v4	modify	Notice
62	CC-BY_v4	modify	modificationNotice
25	CC-BY-NC-SA_v4	derive	Attribution
25	CC-BY-NC-SA_v4	derive	Notice
25	CC-BY-NC-SA_v4	derive	modificationNotice
25	CC-BY-NC-SA_v4	distribute	Attribution
25	CC-BY-NC-SA_v4	distribute	Notice
25	CC-BY-NC-SA_v4	modify	Attribution
25	CC-BY-NC-SA_v4	modify	Notice
25	CC-BY-NC-SA_v4	modify	modificationNotice
16	CC-BY-NC_v4	derive	Attribution
16	CC-BY-NC_v4	derive	Notice
16	CC-BY-NC_v4	derive	modificationNotice
16	CC-BY-NC_v4	distribute	Attribution
16	CC-BY-NC_v4	distribute	Notice
16	CC-BY-NC_v4	modify	Attribution
16	CC-BY-NC_v4	modify	Notice
16	CC-BY-NC_v4	modify	modificationNotice
13	CC-BY-SA_v4	derive	Attribution
13	CC-BY-SA_v4	derive	Notice
13	CC-BY-SA_v4	derive	modificationNotice
13	CC-BY-SA_v4	distribute	Attribution
13	CC-BY-SA_v4	distribute	Notice
13	CC-BY-SA_v4	modify	Attribution
13	CC-BY-SA_v4	modify	Notice
13	CC-BY-SA_v4	modify	modificationNotice
13	GNU_GPL_v3	derive	Attribution
13	GNU_GPL_v3	derive	Notice
13	GNU_GPL_v3	derive	SourceCode
13	GNU_GPL_v3	derive	modificationNotice
13	GNU_GPL_v3	distribute	Attribution
13	GNU_GPL_v3	distribute	Notice
13	GNU_GPL_v3	distribute	SourceCode
13	GNU_GPL_v3	modify	Attribution
13	GNU_GPL_v3	modify	Notice
13	GNU_GPL_v3	modify	SourceCode
13	GNU_GPL_v3	modify	modificationNotice
3	MIT	derive	Attribution
3	MIT	derive	Notice
3	MIT	derive	modificationNotice
3	MIT	distribute	Attribution
3	MIT	distribute	Notice
3	MIT	modify	Attribution
3	MIT	modify	Notice
3	MIT	modify	modificationNotice

## 7 Compliance to the FAIR principles

This deliverable concerns an analysis on a meta-level. This means its results influence the metadata structure and content of many Polifonia resources. The only **Data** component delivered is a resulting Knowledge Graph of licences. The workflows (in the experiments) represents the Polifonia REEC (Research Ecosystem) component types **Executable**, **Re-Usable Software** and **Data**.

The experiments are documented in two GitHub repositories, sub-repositories to the Polifonia-project github.

(A) Experiments with musoW and Large Language Models (LLM): <https://github.com/polifonia-project/musow-licences-experiments-llm> (as discussed in Chapter 4)

(b) Project with knowledge graph construction code: <https://github.com/polifonia-project/licences> (as discussed in Chapter 5)

The documentation of software components which are part of these two experiments in GitHub and their partly release in Zenodo ensures the re-usability and evaluation by the Polifonia community (and beyond).

By departing from current recommendations about how to formalise licences, as developed by W3C and in projects such as [FAIRImpact](#), the experiments documented in this deliverable are firmly adhering the FAIR principles in particular those of Re-Use and Interoperability.

To which extent the outcome of these experiments will be turned into formal components in the Polifonia Research Ecosystem needs to be decided in future work. This concerns in particular possible implications for the Polifonia Ontology Network (D2.2.). Results of this ongoing work will be reported in the Final Polifonia Research Data Management Plan (D7.3) which will be ready towards the end of the project (M38).

*List all components related to this DL*

(A) Experiments with musoW and Large Language Models (LLM) –musoW licences from web pages with ChatGPT contains one component:

name	musoW licences from web pages with ChatGPT
component-id	musow-licences
type	Experiment
work-package	WP2
related-components	musoW dataset; Sparql-anything; Dalicc <a href="https://www.dalicc.net/">https://www.dalicc.net/</a> ; OpenAI platform <a href="https://platform.openai.com/">https://platform.openai.com/</a>
licence	CC-BY_v4
contributors	Enrico Daga
link	<a href="https://github.com/polifonia-project/musow-licences-experiments-llm">https://github.com/polifonia-project/musow-licences-experiments-llm</a>
release-version	d2.6-v1.0
doi	<a href="https://doi.org/10.5281/zenodo.10054025">https://doi.org/10.5281/zenodo.10054025</a>

(B) Project with knowledge graph construction code – License Knowledge Graph contains two components: Licences KG generation pipeline and Licences Knowledge Graph:



name	Licences KG generation pipeline
component-id	licences-pipeline
work-package	WP2
type	Software
licence	CC-BY_v4
contributors	Enrico Daga
release-version	d2.6-v1.0
link	<a href="https://github.com/polifonia-project/licences/blob/main/Pipeline.md">https://github.com/polifonia-project/licences/blob/main/Pipeline.md</a>
doi	<a href="https://doi.org/10.5281/zenodo.10054031">https://doi.org/10.5281/zenodo.10054031</a>

name	Licences Knowledge Graph
component-id	licences
type	KnowledgeGraph
work-package	WP2
licence	CC-BY_v4
contributors	Enrico Daga
link	<a href="https://github.com/polifonia-project/licences/blob/main/KG.md">https://github.com/polifonia-project/licences/blob/main/KG.md</a>
doi	<a href="https://doi.org/10.5281/zenodo.10054031">https://doi.org/10.5281/zenodo.10054031</a>

All components adhere to the Polifonia 10 Rules for Open Science [74].

## 8 Conclusions

Ensuring legal compliance in the reuse of musical cultural heritage is crucial for Polifonia researchers and their collaboration with project stakeholders inside and outside academia. In the study reported in this deliverable, we utilized semantic web technologies to integrate concepts, relations, and entities related to licences and terms of use. Our primary focus was on addressing issues related to the exploitation and expansion of music collection metadata to provide answers related to the rights and terms of use of the data. This approach aimed to support stakeholders such as music researchers, computational musicologists, and industry practitioners in selecting and evaluating the appropriateness of musical data sources for their projects.

We departed from an analysis of licences as documented in the musoW catalogue, which contains a variety of metadata information of musical data sources, including those used in Polifonia. The inspection of the existing content of the metadata field on licences revealed the variety of licence information used but also how unstructured this information often is, and that in many cases information is also missing.

From this analysis we departed in the direction to complement the content of this metadata field. To expand the coverage of the musoW metadata to include licence information that contributors may have missed, we explored the application of a Large Language Model (LLM) for linking and extracting licence information from the original web resources, the links to which are also included in the catalogue. In this work, we particularly explored the application of LLMs for extracting licence information. The method is capable of returning one (or more) licences associated to the resource and was proven to be sufficiently reliable on that specific task. In this work, we focused on identification and linkage of licence metadata using LLMs and experimented with one specific language model (the Open AI GPT 3.5).

Furthermore, we worked towards a harmonisation of the information or content of the licence information. Here, we mapped the information found about licences to structured information provided about licences using the ODRL model. The main result here is the Polifonia Licenses Knowledge Graph, which connects resources associated with their terms of use represented computationally.

Last, but not least, we also validated the appropriateness of the ODRL expressed terms of use to answer queries relevant to musicology researchers. Additionally, we verified the suitability of RDF licence catalogues such as Dalicc to cover the range of licence-related questions relevant to musicological research, using the musoW catalogue as a reference data set.

The method(s) we applied to extract licence information can be used to extract other IP-related information, such as copyright and terms of use (information types which we left to future work). One potential interesting research direction relates to the extraction of specific terms from the text, for example, whether the web page explicitly states that the provided content cannot be used for commercial purposes (without specifying a specific licence).

Concerning the details of the method, as described above we used *one* specific LLM. Future work needs to include comparing the results of these experiments with other LLMs, with different styles of prompts, in order to study the sensitivity of the behaviour in relation to prompt variation.

As we argued for throughout this deliverable, licence information is a critical piece of metadata in data catalogues and errors in the information given about licences can have negative impacts. It can hinder - unnecessarily - the re-use (if information is missing), or it can entice misuse. Extending catalogue metadata with licence information will help in mitigating those issues. However, in the future, we plan to integrate the automatic enrichment of metadata content as provenance information, explicitly and expressively stating that the licence was automatically generated (and therefore, inviting the user to verify the information). Moreover, as part of a provenance workflow, we could not only use 'living web sources' but also to create permalinks to *frozen*, i.e. archived snapshots of web sources (at the time of the access) as provided for instance by the Wayback machine of the [Internet Archive](#). This way a persistent and immutable reference will be created, and link rot avoided [75, 76].

Finally, we want to integrate our results further within other workflows in the Polifonia project. The outcome of this work leads to better licence information for the components of the Polifonia Research Ecosystem. We will use the Technical Board governance to contact Pilot and engineers/developers working in the various Pilots, to send them the details of their licence choices, and to show the implications of their choices in terms of permissions, prohibitions, and duties. All these actions contribute to make sure that the chosen licence actually meets their true needs and intended information signals. We plan to design a method and tool to further support the annotation of components of the Polifonia Research Ecosystem such as data, software/applications, and other forms of documentation, with computational licences. By embedding such a method firmly into Research Ecosystem Framework <sup>1</sup> currently under development by the consortium, we ensure that can be useful beyond Polifonia. This report (D2.6) needs to be seen in context and related to upcoming reports such as D5.7 in which the current statuts of the Polifonia Ecosystem, will be presented (D5.7 Tool support for citing and annotating musical scholarly objects) and D7.3 (Final data management plan).

---

<sup>1</sup>Research Ecosystem Framework (REECO): <http://github.com/reeco-framework>

## Bibliography

- [1] S. Chambers, G. Palkó, F. Morselli, K. Ferguson, and A. Scharnhorst, *Book of Abstracts, DARIAH Annual Event 2023: Cultural Heritage Data as Humanities Research Data?* Zenodo, 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.8340671>
- [2] G. Governatori, A. Rotolo, S. Villata, and F. Gandon, “One License to Compose Them All,” in *The Semantic Web—ISWC 2013*. Springer, 2013, pp. 151–166.
- [3] C. Cardellino, S. Villata, F. Gandon, G. Governatori, B. Lam, and A. Rotolo, “Licentia: a Tool for Supporting Users in Data Licensing on the Web of Data,” in *Proceedings of the ISWC 2014 Posters & Demonstrations Track, a track within the 13th International Semantic Web Conference (ISWC 2014)*, M. Horridge, M. Rospocher, and J. van Ossenbruggen, Eds., Riva del Garda, Italy, 21 October 2014.
- [4] E. Daga, M. d’Aquin, A. Gangemi, and E. Motta, “Propagation of Policies in Rich Data Flows,” in *Proceedings of the 8th International Conference on Knowledge Capture*. ACM, 2015, p. 5.
- [5] T. Pellegrini, G. Havur, S. Steyskal, O. Panasiuk, A. Fensel, V. Mireles, T. Thurner, A. Polleres, S. Kirrane, and A. Schönhofer, “Dalicc: a license management framework for digital assets,” *Proceedings of the Internationales Rechtsinformatik Symposium (IRIS)*, vol. 10, 2019.
- [6] M. Daquino, D. Enrico, D. Mathieu, A. Gangemi, H. Simon, L. Robin, A. MERONO PENUJELA, M. Paul *et al.*, “Characterizing the landscape of musical data on the web: State of the art and challenges,” in *WHiSe 2017 Workshop on Humanities in the Semantic Web Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017)*, 2017, pp. 57–68.
- [7] T. Bottini, C. Guillotel-Nothmann, M. Gurrieri, and F. Pouillet-Pagès, “Tonalties: a collaborative annotation interface for music analysis,” in *Musical Heritage Knowledge Graphs workshop during the 22nd International Semantic Web Conference 2022*. RWTH Aachen University., 2022.
- [8] B. L. Sturm, “Concatenative sound synthesis and intellectual property: An analysis of the legal issues surrounding the synthesis of novel sounds from copyright-protected work,” *Journal of New Music Research*, vol. 35, no. 1, pp. 23–33, 2006.
- [9] J. Grobelny, “Mashups, sampling, and authorship: A mashupsampliography,” *Music Reference Services Quarterly*, vol. 11, no. 3-4, pp. 229–239, 2007.
- [10] J. Delgado, I. Gallego, and E. Rodríguez, “Use of a rights expression language for secure music distribution,” *Journal of New Music Research*, vol. 34, no. 2, pp. 185–195, 2005.
- [11] L. Kagal, T. Finin, and A. Joshi, “A policy based approach to security for the semantic web,” in *International Semantic Web Conference*. Springer, 2003, pp. 402–418.
- [12] C. Flavián and M. Guinalú, “Consumer trust, perceived security and privacy policy: three basic elements of loyalty to a web site,” *Industrial Management & Data Systems*, vol. 106, no. 5, pp. 601–620, 2006.
- [13] D. Scott and R. Sharp, “Abstracting application-level web security,” in *Proceedings of the 11th international conference on World Wide Web*. ACM, 2002, pp. 396–407.
- [14] Y. Qu, X. Zhang, and H. Li, “OREL: an ontology-based rights expression language,” in *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, 2004, pp. 324–325.
- [15] E. Yuan and J. Tong, “Attributed based access control (ABAC) for web services,” in *Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on*. IEEE, 2005.
- [16] G. Tonti, J. M. Bradshaw, R. Jeffers, R. Montanari, N. Suri, and A. Uszok, “Semantic Web languages for policy representation and reasoning: A comparison of KAOs, Rei, and Ponder,” in *International Semantic Web Conference*. Springer, 2003, pp. 419–437.

- [17] R. Iannella, "Open digital rights management," in *World Wide Web Consortium (W3C) DRM Workshop*, 2001.
- [18] V. R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, "Law and the semantic web, an introduction," in *Law and the Semantic Web*. Springer, 2005, pp. 1–17.
- [19] R. Braukmann, W. Hugo, and J. van Kemenade, "Harmonising Access Procedures for Sensitive Data - Workshop at the Open Science Conference," Jul. 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.8108414>
- [20] P. N. Gray, V. Brookfield, and S. M. McJohn, "Artificial Legal Intelligence," *Harvard Journal of Law & Technology*, vol. 12, no. 1, 1998.
- [21] O. Corcho, M. Fernández-López, A. Gómez-Pérez, and A. López-Cima, "Building legal ontologies with METHONTOLOGY and WebODE," in *Law and the semantic web*. Springer, 2005, pp. 142–157.
- [22] A. Gangemi, M.-T. Sagri, and D. Tiscornia, "A constructive framework for legal ontologies," in *Law and the semantic web*. Springer, 2005, pp. 97–124.
- [23] B. Alexander, "LKIF core: Principled ontology development for the legal domain," *Law, ontologies and the semantic web: channelling the legal information flood*, vol. 188, p. 21, 2009.
- [24] J. Breuker *et al.*, "Constructing a legal core ontology: LRI-Core," in *Proceedings WONTO-2004, Workshop on ontologies and their applications*. IVI (FNWI), 2004, pp. 115–126.
- [25] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer *et al.*, "The LKIF Core Ontology of Basic Legal Concepts." *LOAIT*, vol. 321, pp. 43–63, 2007.
- [26] A. Gangemi, A. Prisco, M.-T. Sagri, G. Steve, and D. Tiscornia, "Some ontological tools to support legal regulatory compliance, with a case study," in *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 607–620.
- [27] P. Casanovas, "Conceptualisation of rights and meta-rule of law for the web of data," *Journal of Governance and Regulation*, vol. 4, 01 2015.
- [28] P. Casanovas, V. Rodríguez-Doncel, C. Santos, and A. Gómez-Pérez, "A European Framework for Regulating Data and Metadata Markets." in *PrivOn@ ISWC*, 2016.
- [29] P. A. Jamkhedkar and G. L. Heileman, "Rights Expression Languages," *Handbook of Research on Secure Multimedia Distribution*. New York: IGI Books, pp. 1–21, 2009.
- [30] G. Governatori and R. Iannella, "A modelling and reasoning framework for social networks policies," *Enterprise Information Systems*, vol. 5, no. 1, pp. 145–167, 2011.
- [31] B. Carminati, E. Ferrari, and B. M. Thuraisingham, "Using RDF for policy specification and enforcement," in *15th International Workshop on Database and Expert Systems Applications (DEXA 2004), with CD-ROM, 30 August - 3 September 2004, Zaragoza, Spain*, 2004, pp. 163–167. [Online]. Available: <http://doi.ieeecomputersociety.org/10.1109/DEXA.2004.173>
- [32] X. Wang, T. DeMartini, B. Wragg, M. Paramasivam, and C. Barlas, "The MPEG-21 rights expression language and rights data dictionary," *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 408–417, 2005.
- [33] J. Hunter, "Adding multimedia to the semantic web: Building an mpeg-7 ontology," in *Proceedings of the First International Conference on Semantic Web Working*. CEUR-WS. org, 2001, pp. 261–283.
- [34] —, "Enhancing the semantic interoperability of multimedia through a core ontology," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 49–58, 2003.
- [35] R. García, R. Gil, and J. Delgado, "A web ontologies framework for digital rights management," *Artificial Intelligence and Law*, vol. 15, no. 2, pp. 137–154, 2007.
- [36] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 standard," *IEEE Transactions on circuits and systems for video technology*, vol. 11, no. 6, pp. 688–695, 2001.
- [37] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, *COMM: designing a well-founded multimedia ontology for the web*. Springer, 2007.

- [38] R. Iannella, S. Guth, D. Pähler, and A. Kasten, “ODRL: Open Digital Rights Language 2.1,” W3C Community Group, Tech. Rep., 2015. [Online]. Available: <https://www.w3.org/ns/odrl/2/ODRL21>
- [39] M. Steidl, R. Iannella, V. Rodríguez-Doncel, and S. Myles, “ODRL vocabulary & expression 2.2,” W3C, W3C Recommendation, Feb. 2018, <https://www.w3.org/TR/2018/REC-odrl-vocab-20180215/>.
- [40] V. Rodríguez-Doncel, S. Villata, and A. Gómez-Pérez, “A dataset of RDF licenses,” in *Legal Knowledge and Information Systems. JURIX 2014: The Twenty-Seventh Annual Conference.*, R. Hoekstra, Ed. IOS Press, 2014.
- [41] S. Steyskal and A. Polleres, “Towards Formal Semantics for ODRL Policies,” in *International Symposium on Rules and Rule Markup Languages for the Semantic Web*. Springer, 2015, pp. 360–375.
- [42] G. M. Kapitsaki, N. D. Tselikas, and I. E. Foukarakis, “An insight into license tools for open source software systems,” *Journal of Systems and Software*, vol. 102, pp. 72–87, 2015.
- [43] T. Tuunanen, J. Koskinen, and T. Kärkkäinen, “Automated software license analysis,” *Automated Software Engineering*, vol. 16, no. 3-4, pp. 455–490, 2009.
- [44] D. M. German, M. Di Penta, and J. Davies, “Understanding and auditing the licensing of open source software distributions,” in *Program Comprehension (ICPC), 2010 IEEE 18th International Conference on*. IEEE, 2010, pp. 84–93.
- [45] Y. Manabe, Y. Hayase, and K. Inoue, “Evolutional analysis of licenses in FOSS,” in *Proceedings of the Joint ERCIM Workshop on Software Evolution (EVOL) and International Workshop on Principles of Software Evolution (IWPSE)*. ACM, 2010, pp. 83–87.
- [46] D. M. German, Y. Manabe, and K. Inoue, “A sentence-matching method for automatic license identification of source code files,” in *Proceedings of the IEEE/ACM international conference on Automated software engineering*. ACM, 2010, pp. 437–446.
- [47] M. Di Penta, D. M. German, and G. Antoniol, “Identifying licensing of jar archives using a code-search approach,” in *Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on*. IEEE, 2010, pp. 151–160.
- [48] I. E. Foukarakis, G. M. Kapitsaki, and N. D. Tselikas, “Choosing Licenses In Free Open Source Software.” in *SEKE*, 2012, pp. 200–204.
- [49] T. A. Alspaugh, H. U. Asuncion, and W. Scacchi, “Analyzing software licenses in open architecture software systems,” in *Proceedings of the 2009 ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development*. IEEE Computer Society, 2009, pp. 54–57.
- [50] T. F. Gordon, “Analyzing open source license compatibility issues with Carneades,” in *Proceedings of the 13th International Conference on Artificial Intelligence and Law*. ACM, 2011, pp. 51–55.
- [51] R. Gavriiloaie, W. Nejdl, D. Olmedilla, K. E. Seamons, and M. Winslett, “No Registration Needed: How to Use Declarative Policies and Negotiation to Access Sensitive Resources on the Semantic Web,” in *The Semantic Web: Research and Applications*. Springer, 2004, pp. 342–356.
- [52] H. Li, X. Zhang, H. Wu, and Y. Qu, “Design and application of rule based access control policies,” in *Proc of the Semantic Web and Policy Workshop*, 2005, pp. 34–41.
- [53] P. A. Bonatti and D. Olmedilla, “Rule-based Policy Representation and Reasoning for the Semantic Web,” in *Proceedings of the Third International Summer School Conference on Reasoning Web*, ser. RW’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 240–268. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2391482.2391488>
- [54] P. Bellini, L. Bertocci, F. Betti, and P. Nesi, “Rights enforcement and licensing understanding for RDF stores aggregating open and private data sets,” in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–6.
- [55] R. Pucella and V. Weissman, “A formal foundation for ODRL,” *arXiv preprint cs/0601085*, 2006.
- [56] M. Sensoy, T. J. Norman, W. W. Vasconcelos, and K. Sycara, “OWL-POLAR: A framework for semantic policy representation and reasoning,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 12, pp. 148–160, 2012.

- [57] S. Villata and F. Gandon, "Licenses compatibility and composition in the web of data," in *Proceedings of the Third International Conference on Consuming Linked Data-Volume 905*. CEUR-WS. org, 2012, pp. 124–135.
- [58] A. Rotolo, S. Villata, and F. Gandon, "A deontic logic semantics for licenses composition in the web of data," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. ACM, 2013, pp. 111–120.
- [59] H.-P. Lam and G. Governatori, "The making of SPINdle," in *Rule Interchange and Applications*. Springer, 2009, pp. 315–322.
- [60] G. Governatori, H.-P. Lam, A. Rotolo, S. Villata, G. Ateazing, and F. Gandon, "Checking licenses compatibility between vocabularies and data," in *Proceedings of the Fifth International Workshop on Consuming Linked Data (COLD2014)*, 2014.
- [61] G. Governatori, H.-P. Lam, A. Rotolo, S. Villata, and F. Gandon, "Heuristics for Licenses Composition." in *JURIX*, 2013, pp. 77–86.
- [62] E. Cabrio, A. Palmero Aproso, and S. Villata, "These Are Your Rights," in *The Semantic Web: Trends and Challenges*, ser. Lecture Notes in Computer Science, V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, Eds. Springer International Publishing, 2014, vol. 8465, pp. 255–269. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-07443-6\\_18](http://dx.doi.org/10.1007/978-3-319-07443-6_18)
- [63] D. N. Ribeiro and K. Forbus, "Combining analogy with language models for knowledge extraction," in *3rd Conference on Automated Knowledge Base Construction*, 2021.
- [64] Z. Yao, Y. Cao, Z. Yang, V. Deshpande, and H. Yu, "Extracting biomedical factual knowledge using pretrained language model and electronic health record context," in *AMIA Annual Symposium Proceedings*, vol. 2022. American Medical Informatics Association, 2022, p. 1188.
- [65] Y. Gu, S. Zhang, N. Usuyama, Y. Woldesenbet, C. Wong, P. Sanapathi, M. Wei, N. Valluri, E. Strandberg, T. Naumann *et al.*, "Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events," *arXiv preprint arXiv:2307.06439*, 2023.
- [66] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," *arXiv preprint arXiv:2102.02503*, 2021.
- [67] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *arXiv preprint arXiv:2306.08302*, 2023.
- [68] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling," *arXiv preprint arXiv:2306.11489*, 2023.
- [69] J. Frey, L.-P. Meyer, N. Arndt, F. Brei, and K. Bulert, "Benchmarking the abilities of large language models for rdf knowledge graph creation and comprehension: How well do llms speak turtle?" *arXiv preprint arXiv:2309.17122*, 2023.
- [70] J. de Berardinis, A. Meroño-Peñuela, A. Poltronieri, and V. Presutti, "Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs," *Scientific Data*, vol. 10, no. 1, p. 641, 2023.
- [71] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, and A. Jain, "Structured information extraction from complex scientific text with fine-tuned large language models," *arXiv preprint arXiv:2212.05238*, 2022.
- [72] E. Daga, L. Asprino, A. Gangemi, and P. Mulholland, "Knowledge Graph Construction with a façade: a unified method to access heterogeneous data sources on the Web," *Transactions on Internet Technologies*, 2023.
- [73] E. Daga, luigi asprino, and J. Dowdy, "Sparql-anything/sparql.anything: v0.8.2," May 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7963150>
- [74] A. Scharnhorst, R. Van Horik, E. Daga, M. Daquino, E. Musumeci, P. Van Kranenburg, C. Guillotel-Nothmann, M. Gurrieri, V. Presutti, M. Clementi, A. Meroño Peñuela, M. Turci, E. Marzi, A. Puglisi, and R. Fournier-S'niehotta, "D7.2 data management plan (second version)," Feb. 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7660299>

- [75] M. Klein, “The why and how of robustifying your links [slides],” Tech. Rep., oct 2022. [Online]. Available: <https://doi.org/10.2172%2F1896395>
- [76] S. M. Jones, M. Klein, and H. van de Sompel, “Robustifying links to combat reference rot,” WorkingPaper, Feb. 2021. [Online]. Available: <https://journal.code4lib.org/articles/15509>