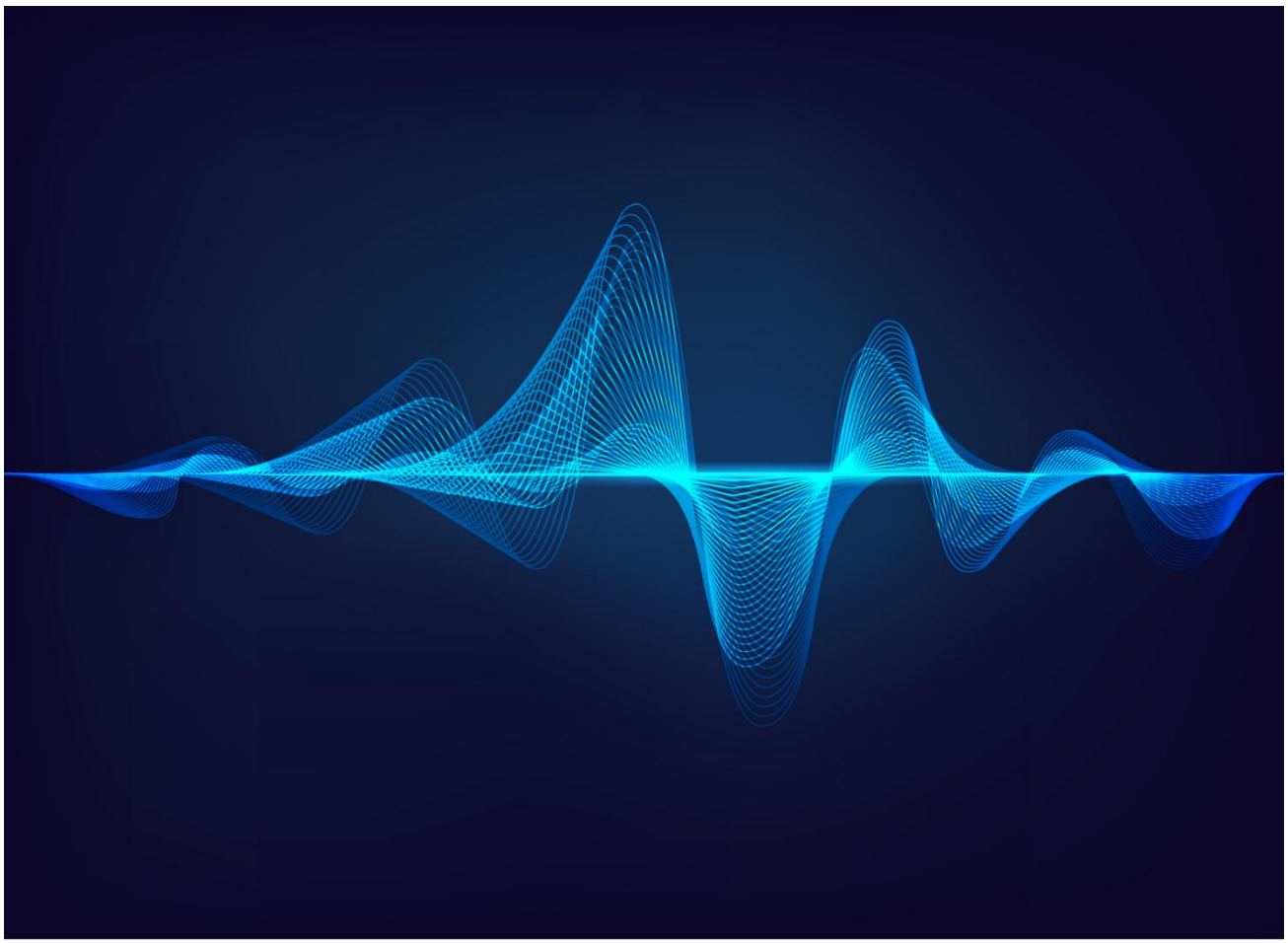


D5.1 – Open Music Observatory

Building an open data sharing space for the European music sector

OpenMusE

An open, scalable data to-policy pipeline for European music ecosystems

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Project Factsheet

Acronym:	**OpenMusE**
Title:	**An open, scalable data to-policy pipeline for European music ecosystems**
Coordinator:	SINUS Markt- und Sozialforschung GmbH
Reference:	101095295
Type:	Research and Innovation Action
Program:	Horizon Europe
Start:	1st January 2023
Duration:	36 months
Website:	https://www.openmuse.eu/
Consortium:	**SINUS Markt- und Sozialforschung GmbH,** Germany (SINUS), Coordinator **TURUN YLIOPISTO,** Finland (UTU) **UNIVERSITEIT VAN AMSTERDAM,** Netherlands (UVA) **Scuola Superiore di Studi Universitari e di Perfezionament,** Italy (SSSA) **EKONOMICKA UNIVERZITA V BRATISLAVE,** Slovakia (EUBA) **Reprex B.V.,** Netherlands (REPREX) **SYNYO GmbH,** Austria (SYNYO) **MUSIC INNOVATION HUB SPA IMPRESA SOCIALE,** Italy (MIH) **Slovenský ochranný zväz autorský pre práva k hudobným,** Slovakia (SOZA) **Aloaded AB,** Sweden (ALOADED) **Music Export Ukraine,** Ukraine (MEU) **Muzikos Eksporto Fondas,** Lithuania (MXF) **ARTISJUS MAGYAR SZERZOI JOGVEDO IRODA EGYESULET,** Hungary (ARTISJUS) **MUSICAUTOR SDRUZHENIE,** Bulgaria (MUSICAUTOR) **HEARDIS! GMBH,** Germany (HEARDIS)

© 2024 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295

Deliverable Factsheet

Number:	**D5.1**
Title:	**Open Music Observatory**
Lead beneficiary:	Reprex
Work package:	WP5
Task:	T5.1
Dissemination level:	PU
Submission date:	2024-05-31
Main author(s):	Daniel Antal, Reprex
Contributor(s):	Andor Kornél Barát, Reprex
	Ádám Lázár, Reprex
Quality check:	István Burus, Reprex
	Felix Haaksman, HearDis!
	Stefan Kragh, Aloaded

Document history (the versions represent the documentation of deliverable, not the version of the information system built, which is not versioned):

Revision	Date	Main modification	Author
0.1	27/05/2024	Draft	Daniel Antal
0.2	29/05/2024	Modification after feedback	Felix Haaksman
0.3	29/05/2024	Modifications after feedback	István Burus
0.9	30/05/2024	Final document	Daniel Antal
1.0	31/05/2024	Format check, submission to EC	Richard Wages
1.1	06/05/2026	Corrections requested by reviewers	Daniel Antal
2.0	10/08/2026	Corrections requested by reviewers	James Edwards

Disclaimer of Warranties

This project has received funding from the European Union's Horizon Europe, research and innovation programme, under Grant Agreement No. 101095295.

This document has been prepared by OpenMusE project partners as an account of work carried out within the framework of the EC-GA contract no 101095295.

Any dissemination of results must indicate that it reflects only the author's view and that the Commission Agency is not responsible for any use that may be made of the information it contains.

Neither Project Coordinator, nor any signatory party of OpenMusE Project Consortium Agreement, nor any person acting on behalf of any of them:

(a) makes any warranty or representation whatsoever, express or implied,
   (i). with respect to the use of any information, apparatus, method, process, or similar item disclosed in this document, including merchantability and fitness for a particular purpose, or
   (ii). that such use does not infringe on or interfere with privately owned rights, including any party's intellectual property, or
   (iii). that this document is suitable to any particular user's circumstance; or

(b) assumes responsibility for any damages or other liability whatsoever (including any consequential damages, even if Project Coordinator or any representative of a signatory party of the OpenMusE Project Consortium Agreement, has been advised of the possibility of such damages) resulting from your selection or use of this document or any information, apparatus, method, process, or similar item disclosed in this document.

Table of contents

Table of contents 5
Glossary 8
Music terms 8
Data science terms 9
Artificial intelligence terms and definitions 11
Abbreviations 11
Executive Summary 14
1 Introduction 16
2 Background & Concept 18
2.1 History of the Music Observatory 18
2.2 Open Music Dataspace 20
2.2.1 European Interoperability Framework 21
2.2.2 Embrace open 22
2.3 Prototyping in Slovakia 23
2.4 Other Observatories 24
2.4.1 Milk market observatory 24
2.4.2 European Audiovisual Observatory (EAO) 25
2.4.3 European Observatory on Infringements of Intellectual Property Rights (EUIPO) 26
2.4.4 European Market Observatory for Fisheries and Aquaculture Products (EUMOFA) 26
3 Core Services 28
3.1 Collect: Data Curation & Collection 30
3.1.1 Microdata, Collections, Records 31
3.1.2 Primary data collection 32
3.1.3 Metadata 33
3.1.4 Statistical indicators and datasets 34
3.2 Process 34
3.2.1 Processing & re-processing microdata 34
3.2.2 Documentation 34
3.3 Disseminate 35
3.3.1 EU Open Data Portal 35
3.3.2 Europeana 36
3.3.3 Wikibase Cloud 37
3.3.4 Music Observatory Website 38
3.3.5 API Endpoint 39

4 Data Catalogue 40
4.1 Collection Guidelines 41
4.2 Topical Pillars 42
4.2.1 Music Economy 43
4.2.2 Music Diversity 45
4.2.3 Music Society & Sustainability 46
4.2.4 Innovation 46
5 Data Sources 47
5.1 Curation of Reusable Data 47
5.1.1 Curation from data vendors 49
5.1.2 Novel data assets 49
5.2 Data Providers 49
6 Standardisation of Data & Terminology 51
6.1 Business Processes 51
6.2 Conceptual and Information Models 51
6.3 Identification & Entity Linking 53
6.3.1 Registers & Authority Files 54
6.3.2 Open and persistent identifiers 55
6.3.3 Not open, music-industry specific identifiers 56
6.3.4 Lyrics 57
6.3.5 ISCC 57
6.3.6 OMO Identifiers 57
7 Data Improvement & Innovation 59
7.1 Value-Added Data Services 59
7.1.1 Data Sharing 59
7.1.2 Fix-the-data 59
7.1.3 Data Linking 59
7.1.4 Registration services 60
7.2 Use Cases 61
7.2.1 Data Health Services for Collective Management 61
7.2.2 Sustainability Reporting for Music Organisations 62
7.2.3 Listen Local 64
7.2.4 Unlabel 65
7.3 Use of AI Systems 66
8 Conclusions & Next Steps: Towards a European Music Observatory 68

8.1 Co-Creating a Governance Model for the Observatory 68
8.1.1 Observatory Stakeholder Network 69
8.1.2 Open Music Data Exchange 70
8.2 Bottom-Up Expansion of the Observatory in a Federation Model 70
8.3 Practical Next Steps in the Project 71
References 74
ANNEXES 78
Annex 1 – Stakeholder profile data sheet for the Observatory Stakeholder Network 78
Annex 2 – Data Curators Manuals 80
Annex 3 – Terminology 85

Glossary

Music terms

Music terms and definitions

Term	Description
audio recording	fixation of sounds [ISRC standard]
author	The author of music or lyrics. Entitled to authors' right or copyright.
composer	The author of a musical work. Entitled to authors' right or copyright.
lyricist	The author of the text of a musical work, or a literary text that is arranged together with a musical work. Entitled to authors' right or copyright.
music video recording	fixation of sounds synchronised with pictures or moving pictures where (a) the fixed sounds are wholly or substantially a musical performance or (b) the *recording* (3.3) is intended for viewing in association with a recording of a musical performance. [ISRC standard]
DSP or digital streaming platform	Digital service providers (DSPs), or Digital Streaming Platforms are companies or organisations that provide access to services online. DSPs can provide access to music downloads, like Apple's iTunes Store, or access to streaming music like Spotify, or even provide satellite-delivered content such as SiriusXM in the USA.
performer	The performer of a musical work; in case of a sound recording, the performer whose performance is fixed in the recording. They may be entitled to neighbouring or sound recording copyrights.
producer	The person or legal entity that produces the recorded fixation of the sound recording. They are entitled to neighbouring or sound recording copyrights.
rights management (organisations)	the function of managing the rights on behalf of rights owners. It can be companies whose sole purpose is to ensure that content that has been licensed has delivered royalties that are identified and accounted for. The role can be taken by collective management organisations or by private companies on behalf of songwriters, composers, performers, music publishers, or record labels.

Data science terms

Data and information science terms, definitions

Term	Description
conceptualisation	an abstract, simplified view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them
data	reinterpretable representation of information in a formalised manner suitable for communication, interpretation, or processing Note 1 to entry: Data can be processed by humans or by automatic means.[SOURCE:ISO/IEC 2382:2015, 2121272]
database	collection of data organised according to a conceptual structure describing the characteristics of these data and the relationships among their corresponding entities, supporting one or more application areas. [SOURCE:ISO/IEC 2382:2015, 2121413]
data set or dataset	identifiable collection of data available for access or download in one or more formats [SOURCE:Adapted from ISO 19115-2:2009, 4.7] *Beware: various conceptual and information models use different dataset definitionns*.
register	an official list or record of names or items; it aims to be a complete list of the objects in a specific group of objects or population, for example, all copyright-protected musical works in a country, or all legal person enterprises in another country.
data flow definition	a structure which describes, categorises and constrains the allowable content of a data set that providers will supply for different reference periods. [SDMX 3.0]
datacube	A statistical data set created in a multi-dimensional space (e.g., time, geography, gender), or hyper-cube, indexed by those dimensions. *The term cube shouldn't be taken literally, it is not meant to imply that there are exactly three dimensions.*
knowledge base or K-base	database that contains inference rules and information about human experience and expertise in a domain. 1: In self-improving systems, the knowledge base additionally contains information resulting from the solution of previously encountered problems. The terms knowledge base and K-base are standardised by ISO/IEC [ISO/IEC 2382-1:1993].
knowledge representation	process or result of encoding and storing knowledge in a knowledge base. Term and definition standardised by ISO/IEC [ISO/IEC 2382-28:1995].
knowledge graph	a knowledge representation that uses a graph-structured data model to represent and operate on data.

Term	Description
knowledge source	source of information from which a knowledge base has been created for a specific kind of problem Term and definition standardised by ISO/IEC [ISO/IEC 2382-28:1995].
knowledge engineering tool	functional tool designed to facilitate the rapid development of knowledge-based systems. 1. A knowledge engineering tool incorporates specific strategies for knowledge representation, inference, and control, as well as elementary modeling constructs for easy handling of typical problems.Term and definition standardised by ISO/IEC [ISO/IEC 2382-28:1995].
conceptual model	representation of the characteristics of a universe of discourse by means of entities and entity relationships (ISO/IEC 2382-17:1999). *In this document, we use conceptual models for models that can be used by humans and computers, too, and we use the information model term for use in IT systems.*
interoperability	Ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged. [SOURCE:ISO/IEC 19941:2017]
data portability	Ability to easily transfer data from one system to another without being required to re-enter data.
metadata	data that define and describe other data [ISO/IEC 11179-1:2023]; we use the more functional definition “a statement about a potentially informative object.”
NERD	Named-entity recognition and disambiguation
persistent identifier	A persistent identifier (or *permanent Identifier* or *handle*), is one that never changes, so that your bookmarks and links don’t break when a website or a database or an API service gets updated.
CIDOC-CRM	The conceptual model of CIDOC, the standard conceptualisation of collection management systems in heritage organisations.
RiC	*Records in Context*, a new conceptual model that replaces the four most important international archiving standards.
Wikibase	Wikibase is a software system that help the collaborative management of knowledge in a central repository. It was originally developed for the management of Wikidata, but it is available now for the creation of private, or public-private partnership knowledge graphs. It is developed by Wikimedia Deutschland.

Artificial intelligence terms and definitions

Term	Description
expert system	knowledge-based system that provides for solving problems in a particular domain or application area by drawing inferences from a knowledge base developed from human expertise 1: The term “expert system” is sometimes used synonymously with “knowledge-based system” but should be taken to emphasise expert knowledge. 2. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems. *Expert systems fall under the definition of the AI Act*. [SOURCE:ISO-IEC-2382-28 * 1995]

Abbreviations

Term	Description
AEPO-ARTIS	Organisation representing European artists-performers. Regroups most of the European CMO representing performers.
CISAC	The International Confederation of Societies of Authors and Composers is an international non-governmental, not-for-profit organisation that aims to protect the rights and promote the interests of creators worldwide.
CSRD	The Corporate Sustainability Reporting Directive (CSRD) is European Union (EU) legislation, effective from 5 January 2023, that requires EU businesses—including qualifying EU subsidiaries of non-EU companies—to disclose their environmental and social impacts, and how their environmental, social and governance (ESG) actions affect their business.
DOI	A DOI (Digital Object Identifier) is a standardised unique number given to many (but not all) articles, papers, & books, by some publishers, to identify a particular publication.
DDI	The Data Documentation Initiative is originating for the world of social sciences data archives and more and more in use in statistical organisations for the documentation of microdata.
DSP	Digital service providers (DSPs), or Digital Streaming Platforms are companies or organisations that provide access to services online.
ECCCH	The European Cultural Heritage Cloud is a European Union initiative for a digital infrastructure that will connect cultural heritage institutions and professionals across the EU.
EFRAG	The European Financial Reporting Advisory Group is a private association established in 2001 with the encouragement of the European Commission to serve the public interest. EFRAG extended its mission in 2022 following the new role assigned to EFRAG in the CSRD, providing Technical Advice to the European Commission in the form of fully

Term	Description
	prepared draft EU Sustainability Reporting Standards and/or draft amendments to these Standards.
EMO	The European Music Observatory (EMO) is envisioned as a hub for collecting and analyzing data on the music sector across Europe. Its primary aim is to address the current gaps and inconsistencies in music data collection, which have been a significant challenge for the sector.
ESRS	The European Sustainability Reporting Standards (ESRS) are a set of guidelines developed by the European Financial Reporting Advisory Group (EFRAG) to standardise sustainability reporting across the European Union. These standards are designed to align with the Corporate Sustainability Reporting Directive (CSRD), which mandates detailed corporate reporting on environmental, social, and governance (ESG) issues for many companies operating within the EUm
GESAC	GESAC comprises together 32 European authors' societies in music, audiovisual, visual arts, literature and drama.
GESIS	Leibniz Institute for the Social Sciences.
GSBPM	The Generic Statistical Business Process Model is a international standard model that "describes and defines the set of business processes needed to produce official statistics."
GSIM	Generic Statistical Information Model: a common abstract representation of data objects manipulated in official statistical production and elaborated as an overarching model for implementation standards such as SDMX or DDI.
ICMP	ICMP is the global trade body representing the music publishing industry worldwide.
ISBN-10	The older format of the ISBN book identifier
ISCC	The International Standard Content Code (ISCC) is an identifier for numerous types of digital assets.
ISNI	the ISO certified global standard number for identifying the millions of contributors to creative works and those active in their distribution.
ISMN	The International Standard Music Number identifies published music sheets.
ISRC	The International Standard Recording Code (ISRC) is the international identification system for sound recordings and music video recordings. It was codified in the ISO 3901 standard in 1986.
ISWN	

Term	Description
ISWC	International Standard Musical Work Code is a unique identifier for musical works. It is adopted as international standard ISO 15707.
ORCID iD	A unique, persistent identifier free of charge to researchers.
PPP	A Public-Private Partnership (PPP) is a collaborative arrangement between government entities and private sector companies aimed at financing, designing, implementing, and operating projects or services traditionally provided by the public sector.
SCAPR	International association for the development of the practical cooperation between performers' collective management organisations (CMOs)
SDMX	Statistical Data and Metadata eXchange (SDMX), is an international initiative that aims at standardising and modernising ("industrialising") the mechanisms and processes for the exchange of statistical data and metadata among international organisations and their member countries.
URI	A Uniform Resource Identifier (URI) is a string of characters used to identify a resource on the internet. This resource can be either abstract or physical, such as a website, an email address, or a file. URIs are essential for enabling interactions with resources over a network using specific protocols.
VIAF	The Virtual International Authority File (VIAF) is an international service that consolidates multiple name authority files into a single database. Their primary goal is to enhance the efficiency and usability of library authority files by linking and merging widely used authority records and making them accessible online.
W3C	The World Wide Web Consortium (W3C) is an international community that develops standards for the World Wide Web. Their mission is to lead the Web to its full potential by creating technical specifications and guidelines that are designed to be open and royalty-free. These standards include HTML, CSS, and other web technologies, which ensure that web content is accessible across different browsers and devices.

Our glossary is harmonised with relevant music-sector specific standards (referred to in Chapter 6) and with the ISO *Information technology — Vocabulary* (ISO 2023); *Information technology — Cloud computing — Taxonomy based data handling for cloud services* (ISO 2020) and *Information technology — Cloud computing — Interoperability and portability* (ISO 2017a).

Executive Summary

Our ambition with the development of the Open Music Observatory is to provide the technological basis and a practical roadmap for creating a *European Music Observatory* in a bottom-up, decentralised way. Instead of waiting for a grand, central agreement on what should a European music observatory be collecting and who should control it, we suggest a pragmatic approach: allow any data owners and collectors who satisfy certain quality and cooperation rules to add their data to an Open Music Observatory; when it reaches a sufficient maturity for use in Europe, then decide if its maintenance requires a new institutional form or not.

Creating the Open Music Observatory is a cornerstone task of the OpenMusE project. This task is running till the end of the project (31 December 2025) with the collection, processing, and dissemination of more data and providing innovative, new data services in line with our exploitation pathways. This report is an accompanying document for the creation of Open Music Observatory as a digital infrastructure on the World Wide Web.

The Open Music Observatory is a digital service provider for the music industry that follows the European Interoperability Framework (EIF) definition for such services with a unique governance model. The governance model and the digital service infrastructure represent a unique innovation that considers many good examples from the European Union and other industries.

An observatory has traditionally been a permanent location for observing terrestrial, marine, or celestial events. In the past 30 years, it has also been used for long-term digital data collection programs for markets, social sciences, and humanities. Our milestone requires the start of this observatory after a lengthy and intensive planning and prototyping phase. It can be seen as a modern reimagination of the data observatory model, or the observatory 2.0. We created a new observatory model that fully aligns with the *European Interoperability Framework* but extends the governance of the digital services beyond public bodies, and allows the creation of a public-private partnership to manage the observatory.

We were informed and influenced by the creation of Europeana (which started out from a similar collaborative project) and their new plans to extend their digital services into the European Cultural Heritage Cloud (ECCH). We aim fully interoperability with Europeana and ECCH, but we also bring a new element into their thinking. While they are mainly aggregating the work of public sector memory institutions, we are building a governance model that allows a more successful cooperation among the private sector and the public music sector.

The Open Music Observatory furthermore hosts many many intelligent data improvement technologies and fuels innovative applications/services in line with our project's exploitation pathways – which is to say, it can be considered an "observatory 3.0". These services are at different maturity levels, but they could not be brought to a testable MVP without building out the minimal digital infrastructure and governance model at this milestone.

The Open Music Observatory is available at: https://openmusicobservatory.eu/

Policy recommendations that accompany the launch of the Observatory are available in OpenMusE deliverable D2.3 and the OMO Green Paper.

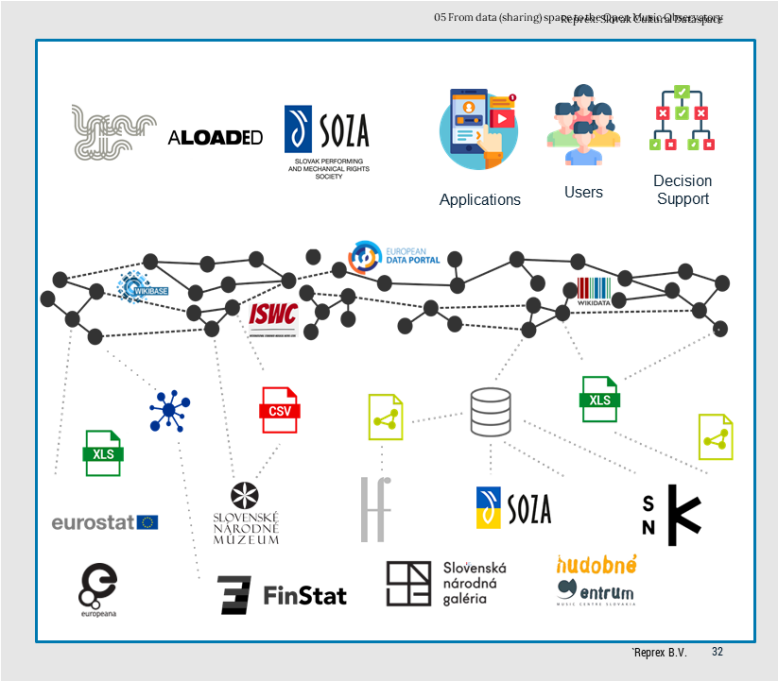
Music Data Space

Application layer: Listen Local Apps for radio or educational playlist editing. Name Entity Matching and Name Entity Disambiguation for rights management.

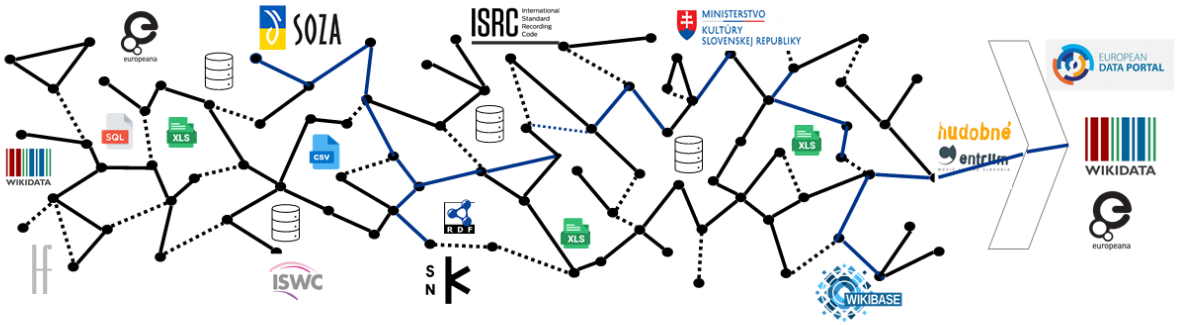
Semantic layer: data coordination among public, private, local and international data owners.

Database layer: databases of participating organisations, shared only as needed and as permitted.

Collection layer: data collection internally (royalty accounts, work registers) and external data sources.



Data sharing and exchange in the Slovak music data sharing space



1. Data enichment

Reprex consolidates the data Excel, CSV, SQL and other database (data file) formats of SOZA, Slovak Music Center, Music Fund, and other organisations into a graph format. We assist curators to find and improve errors in their metadata. We ensure that the resulting data is more usable for rights management, heritage management, publishing.

2. Data dissemination

We send the enriched and proprietary (confidentia) data to the systems of SOZA, Slovak Music Center, Music Fund or other participants. We sent the public data to the EU Open Data Portal (statistical data), to EOSC (data of scientific value), collections data (Europeana and ECCH), and biographical and repertoire data to Wikidata and Wikpeida. This way streamng providers can use reliable data about Slovak music. Reprex B.V. 9

This document is licensed under the CC BY 4.0 LEGAL CODE Attribution 4.0 International license. You must refer to the document with the DOI 10.5281/zenodo.11385044.

Canonical Licence URL: https://creativecommons.org/licenses/by/4.0/ **Other formats**: Plain Text; RDF/XML| **See the deed**

1 Introduction

Task 5.1 *Adding Economy, Diversity, Innovation and Society, Citizenship Pillars to the MVP of the Open Music Observatory and raising it to at least TLR Level 7* is a cornerstone task of the OpenMusE project. It focuses on the ultimate ambition of the project, focusing on creating and validating the technological, service and governance components of an observatory that can become the basis of a future European Music Observatory.

Deliverable 5.1 *Open Music Observatory* is the most important milestone of this task, the actual release of the Open Music Observatory on the world wide web.

The present report is an accompanying document to the deliverable. Its aim is to explain the workings of the Observatory, the concepts underlying its development, and our plans to further develop it to achieve our ultimate ambition that it can become a basis for a future European Music Observatory.

After the **?@sec-executivesummary** and this introduction (Chapter 1), our report is structured as follows:

Chapter 2 introduces the history of the plans to set up a European music observatory and a decentralised approach, the creation of CEEMID–a background of the Open Music Europe project and an important building block established in the *EMO feasibility study*. We then conceptualise the Open Music Observatory as a data (sharing) space, a novel technical and legal innovation of the European Union. We place an emphasis on compliance with the European Interoperability Framework and embracing the open collaboration method, open-source software and open data as fundamental building blocks. At last, we review a few observatories, including the ones that were reviewed in the *EMO feasibility study* to see what services they offer in 2024, and how our service structure compares to theirs.

Chapter 3 provides an overview of the current, core services of the Observatory, which can be seen as the services of a modern statistical vendor and open data portal. In line with the EIF, we integrate our services into the EU Open Data portal (as our main gateway to disseminate statistical datasets), the Europeana and the European Culture Heritage Cloud (for music collections), and Wikipedia's ecosystem (Wikidata and Wikibase) to disseminate data to open knowledge graphs.

Chapter 4 introduces our digital curation guidelines and the nucleus of our data offering. Prior to the release of Observatory 2.0, we could not start ingesting large and valuable datasets, so the catalogue currently contains datasets only from the background IP of the Open Music Europe project. Our data can already be tried out, and we plan to showcase the first extensive datasets and onboarding case studies at several international conferences in November 2025. One of these large-scale case studies is already underway, and it will be introduced later in Section 3.1.1.

Chapter 5 revisits the EMO feasibility study and provides an updated landscape on data availability and potential data partners in the EU. At the time of writing the feasibility study, many pan-European music organisations promised data to a new observatory. We will introduce them the new Open Data Observatory and will invite them to the Observatory Stakeholder Network and ask to give us at least sample datasets for publication.

Chapter 6 shows the data and terminology standardisation work started by the Open Music Observatory. The feasibility study correctly identified this as a key, missing function. We divide this standardisation into three parts:

- ☒ standardisation of music-related of conceptual and information model related to the data we want to collect and disseminate;
- ☒ a standardisation of entity identities such as author name, works, recordings;
- ☒ standardisation of music-related and cross-domain concepts, such as “average income” or “composer” or “lyricist”.

Chapter 7, which provides an overview of the “Observatory 3.0” concept, presenting some of our planned value-added services that aim to embrace the idea of trustworthy data and use of trustworthy AI for the benefit of the music sector fully. We also briefly introduce the innovative services that that consortium partners are planning with the machine-actionable data our APIs will provide them.

Chapter 8 provides a conclusion along with a technical and governance roadmap to deliver a full-fledged Open Music Observatory, which may be adopted by the relevant stakeholders as a European Music Observatory by the end of our project. It embraces the technology of linked open data, the connection of open and enterprise knowledge graphs, and interoperability system of the data sharing space model. We are piloting this structure in the Slovak Republic, where we will be introducing in November the Slovak Music Datapsace with rich applications. This dataspaces can federated with similar national, genre- or sub-sector specific data sharing building blocks into a truly pan-European music observatory.

Annexes to the report showcase documents that provide further insight to the status and the workings of the Open Music Observatory.

2 Background & Concept

This chapter introduces the history of the plans to set up a European music observatory and a decentralised approach, the creation of CEEMID–a background of the Open Music Europe project and an important building block established in the *EMO feasibility study*.

The key lesson of the CEEMID project was that, unlike observatories that started in the late 20th and early 21st century, modern data ecosystems grow faster and can incorporate big data sources if they are decentralised. Therefore, we conceptualise the Open Music Observatory as a data (sharing) space, a novel technical and legal innovation of the European Union. (See Section 2.2.) We place an emphasis on compliance with the European Interoperability Framework and embracing the open collaboration method, open source software and open data as fundamental building blocks.

At last, we review a few observatories (Section 2.4), including the ones that were reviewed in the *EMO feasibility study* to see what services they offer in 2024, and how our service structure compares to theirs.

2.1 History of the Music Observatory

In late 2015, the European Commission started a dialogue with representatives from the music sector1 in Europe to identify key challenges and possible ways to tackle them, including EU support. *Music Moves Europe* has become the framework for these discussions and, more broadly, for EU initiatives and actions to promote the diversity and competitiveness of Europe's music sector regarding policy and funding. As part of the 2018 Preparatory Action *Music Moves Europe: Boosting European Music Diversity and Talent* the EU commissioned the creation of *The feasibility study for the establishment of a European Music Observatory* (European Commission et al. 2020) (in short: EMO Feasibility Study).

In 2020, *Feasibility Study for the Establishment of a European Music Observatory : Final Report* enumerates 45 data gaps where music stakeholders and policymakers are lacking information to develop business and policy actions to increase the competitiveness, value-added and job creation capacity, and diversity of the European music ecosystem. It also called out the fact that "data collection in Eastern and Southern Europe is lagging in comparison to other European Member States in Northern and Western Europe", and highlighted the importance of the former CEEMID project (a predecessor of *Open Music Europe*) as a possible data source for a new European Music Observatory.

"Data collection in Eastern and Southern Europe is lagging in comparison to other European Member States in Northern and Western Europe, with the respective music sectors in most Eastern Europe countries and smaller EU Member States not fully developed, and lacking the tools and processes to gather economic, cultural and social data on the music sector. These data conditions and the problems they present for effective management and policy development are the fundamental reasons for supporting the creation of a European Music Observatory. In particular, robust and meaningful comparative data collected at a regular basis are essential when it comes to assessing the need for interventions at the EU level to address gaps in the market and enhance the efficiencies and global competitiveness of the sector". (European Commission et al. 2020, p9–10)

It also mentioned the work of CEEMID, a bottom-up initiative originally started by three collective management societies (Artisjus et al. 2014). Our proposal wants to put the former CEEMID, currently

1 See European Commission (2021a) about the policy-making process and objectives.

named Digital Music Observatory, a more solid scientific and methodological foundation to make it more user-friendly and to exploit state-of-art statistical, data science and computer science methods to provide more comprehensive, timely, and accurate services for the European music sector.

The former CEEMID (originally: *Central & Eastern European Music Industry Databases*) collaboration offered data and a more modern, decentralised organisational alternative to the creation of the observatory. CEEMID was organised in 2014 according to the principles of a so-called data (sharing) space, which by 2024 would be a cornerstone of the European data strategy and a recognised organisational and legal vehicle to create interoperable digital services in the EU. Initiated by four organisations (out of which three are members of the Open Music Europe consortium) and eventually joined by more than 60 stakeholders in 12 European countries to fill in some of these gaps with a) voluntary data integration among partners, b) open data re-processing c) co-financed data collection (Antal 2020a).

Three former CEEMID partners took a subjective view of the data gaps identified by the feasibility study and, together with further academic and industry partners, applied for a highly competitive Horizon Europe Research and Innovation Action Grant. After winning the grant, we formed the Open Music Europe Consortium and collected the data.

Since the publication of the *EMO feasibility study*, the fruits of the *European Interoperability Framework* applied to digital public services and the European Open Science Cloud, or the planned European Cultural Heritage Cloud, provide solutions that strongly favour a decentralised information model for a new observatory. We imagine a future European Music Observatory that is not a specialised knowledge institution or a library, archive, museum, or statistical agency. Instead, applying the *European Interoperability Framework* should be able to consolidate knowledge from all such knowledge and statistical institutions and find ways to combine data from private enterprises and data collection programs to fill the information gaps of the European music sector stakeholders.

[…]**Potential stakeholders for the purposes of this project: Industry**: those organisations and agents who are linked to the economy of the music sector, representing commercial, for profit interests only. Example: commercial organisations and companies which are involved in the business of music making including organisations which represent those involved in income-generation from the performance, recording, distribution and creation of music. **Civic**: those organisations and agents who are linked to the policies affecting the music sector. Civic should be organisations with a general interest mandate, professional associations receiving public funding, and/or including public entities in their membership. Example: political and non-political policymakers whose decisions impact on performance, recording, distribution and creation of music including NGOs and funding distributors. **Public**: those organisations and agents who are linked to the wider culture of music making and consumption. Example: organisations representing consumers, voluntary and third sector, education and training sector with an interest in the performance, recording, distribution and creation of music. (European Commission et al. 2020, p30)

2.2 Open Music Dataspace

> Open Music Dataspace: as the data management of the Open Music Observatory, it recognises that in a large-scale integration of music sector data involving potentially thousands of data sources, it is impossible to create a unifying schema across all sources. It is an exchange, processing, sharing and provision of data between trusted partners, for a fee or not. It is not necessarily copying or repatriating data centrally but ensures that each data holder has complete control over the conditions (e.g., who, when, and under what conditions) of access to their data. We create this data-sharing space view that certain national, genre-specific or other interest groups may want to apply a deeper level of integration and exchange and, through federation, extend the part of their data sharing to the entire Open Music Observatory. The application of automatic matching and mapping generation techniques, widely used conceptual models, and the concepts of the European Interoperability Framework allow the Observatory Stakeholder Network to integrate data on an 'as-needed' or as-permitted basis, keeping the data ready for integration whenever the data owners agree on such a need.

Our system definition for the Open Music Observatory, we follow the data (sharing) space model that has been developed in recent years and has become a foundational concept of the digital aspect of the strategic policy agenda of triple transition to a more sustainable, digital and just European society.

The data (sharing) space concept has been evolving in recent years. There are several complementary and consistent definitions available that place more or less emphasis on the business processes, the technical processes, or the domain-specific problems. We have considered the following definitions when we arrived to our Open Music Dataspace definition above.

In *Dataspaces: Fundamentals, Principles, and Technique*: (Curry 2020) defines a dataspace as "an emerging approach to data management which recognises that in large-scale integration scenarios, involving thousands of data sources, it is difficult and expensive to obtain an upfront unifying schema across all sources. Data is integrated on an 'as-needed' basis, with the labour-intensive aspects of data integration postponed until they are required. Dataspaces reduce the initial effort required to set up data integration by relying on automatic matching and mapping generation techniques." This definition stresses prioritising an ongoing, dynamic data curation process instead of the less realistic "one-solution-fits-all" approach applied earlier. Extending this definition in the *Design Principles for Data Spaces* (Position paper) adds the idea of *federation*: "From a technical perspective, a data space can be seen as a data integration concept which does not require common database schemas and physical data integration, but is rather based on distributed data stores and integration on an"as needed" basis on a semantic level. Abstracted from this technical definition, a data space can be defined as a federated data ecosystem within a certain application domain and based on shared policies and rules." (Nagel and Lycklama 2021, p7). The ability to start building similarly organised data-sharing spaces with similar enough principles to federate them into a larger union of data space is a critically important idea in our system and governance design.

`

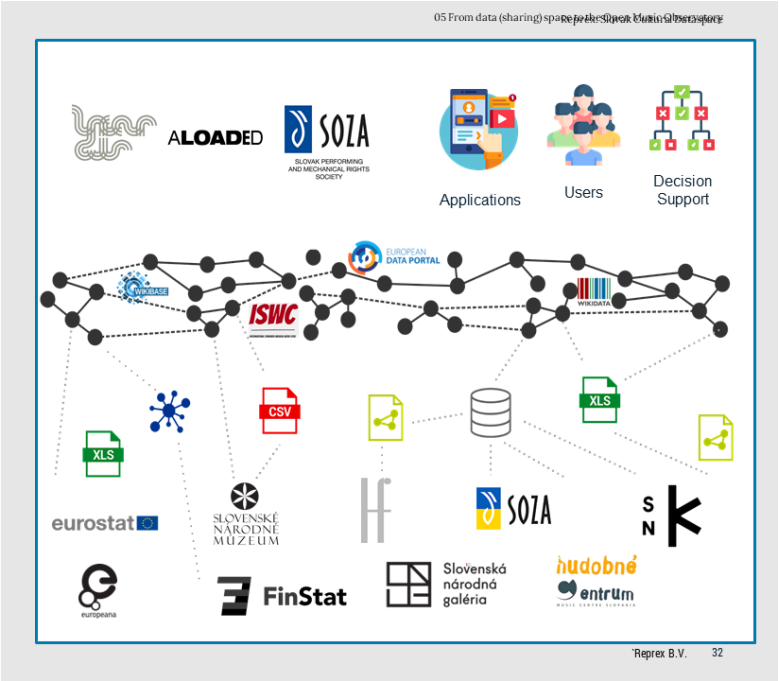
Music Data Space

Application layer: Listen Local Apps for radio or educational playlist editing. Name Entity Matching and Name Entity Disambiguation for rights management.

Semantic layer: data coordination among public, private, local and international data owners.

Database layer: databases of participating organisations, shared only as needed and as permitted.

Collection layer: data collection internally (royalty accounts, work registers) and external data sources.

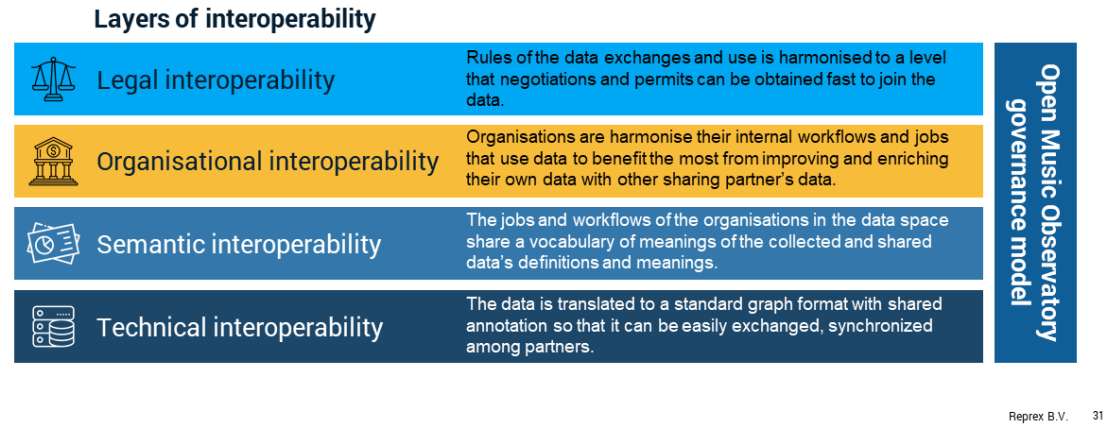


The position paper of the European Broadcasting Union and Gaia-X *Dataspace for Cultural and Creative Industries* highlights the aspect that a "data space is an ecosystem of exchange, processing, sharing and provision of data between trusted partners, for a fee or not. It is not about copying or repatriating data centrally, but about ensuring that each data holder has full control over the conditions (e.g., who, when, and under what condition) of access to their data." (EBU and Gaia-X 2022, p16) Not only because this position paper deals with cultural and creative industries data but also because it stresses the need to provide protocols for reconciliation between open and public and private data sources, we took special attention to this position paper.

2.2.1 European Interoperability Framework

data interoperability: the ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged. [SOURCE:ISO/IEC 19941:2017]

data portability: the ability to easily transfer data from one system to another without being required to re-enter data.



Our innovation is the application of the EIF to a cross-section of public and privately held data, and designing an observatory that is supported by an interoperable data (sharing) space.

2.2.2 Embrace open

According to the study published by the European Commission on the impact of open-source software (OSS) and open-source hardware (OSH) on the European economy, conducted by *Fraunhofer ISI* and *Open Forum Europe* on 6 September 2021, open-source software contributed between €65 to €95 billion to the European Union's GDP. It promises significant growth opportunities for the region's digital economy. The 2020 report on the *Economic Value of Open Data* estimated the value of open data available in the European economy at €184 billion and is forecasted to reach between €199.51 and €334.21 billion in 2025. To unlock this potential, the study makes similar but less specific recommendations as the OSS/OSH study described above2.

2 See *The Impact of Open Source Software and Hardware on Technological Independence, Competitiveness and Innovation in the EU Economy* (European Commission et al. 2021)and *Economic Value of Open Data* (Huyer and van Knippenberg 2020).

2.3 Prototyping in Slovakia

Because a significant part of the Open Music Europe background was developed or tested in Slovakia, we decided to start prototyping the Open Music Observatory in this country.

> Slovak Music Dataspace: a dataspace organised by SOZA, Reprex, and Hudobné centrum to share and exchange data on music related to the Slovak Republic. It provides a secure data exchange supported by trustworthy AI to harmonise and exchange data among representative Slovak music organisations and to create the public Slovak Music Register and the Slovak Comprehensive Music Database.

In a significant stride towards our shared vision, we signed a Memorandum of Understanding in March 2023 (Ministerstvo kultúry SR and Open Music Europe 2023). This strategic alliance includes key stakeholders, such as the Ministry of Culture, and is aimed at establishing a robust public-private partnership. This partnership will play a pivotal role in sharing, exchanging, and improving data related to music in the Slovak Republic, thereby fostering a vibrant music ecosystem.

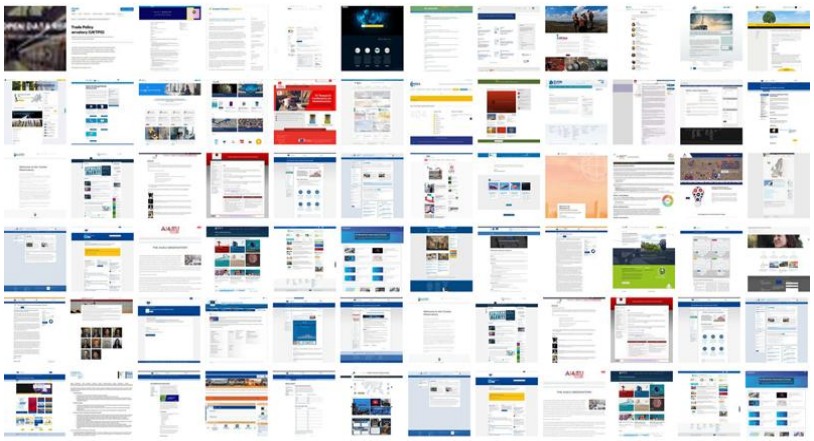
Slovakia has a relatively advanced statistical system: one of the few EU member states with a satellite accounting system for the cultural and creative industry to augment the country's national accounts. The methodological work related to coordinating governmental and privately held data is the subject of other tasks; in making the Open Music Observatory, we only provide the infrastructure and the dissemination of knowledge for replication.



Because we apply GSIM, DDI and SDMX, like the Slovak statistical system, our solutions are replicable in an EU/EEA member state or candidate country that has sufficiently aligned its statistical practices with the European Statistical System.

2.4 Other Observatories

The *EMO feasibility study* presented several observatories to compare their services and organisational models. As more than five years have passed since the data collection for that study, we highlight here some of the considered comparators and add a few more.



Various UN and OECD bodies, and particularly the European Union support or maintain more than 60 data observatories, or permanent data collection and dissemination points.

2.4.1 Milk market observatory

The aim of the EU milk market observatory is to provide the EU dairy sector with more transparency by means of disseminating market data and short-term analysis in a timely manner.



The European milk market observatory celebrates its 10th anniversary [link↗](https://agriculture.ec.europa.eu/news/european-milk-market-observatory-10th-anniversary-2024-04-16_en)

- Publications: regular, in PDF document.
- Statistical data access: Agri-food data portal
- API: Agri-food data API
- EU Open Data portal: yes, via Directorate-General for Agriculture and Rural Development
- License: The Commission's reuse policy is implemented by the Commission Decision of 12 December 2011 on the reuse of Commission documents. Unless otherwise indicated (e.g. in individual copyright notices), content owned by the EU on this website is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. This means that reuse is allowed, provided appropriate credit is given and changes are indicated.

2.4.2 European Audiovisual Observatory (EAO)

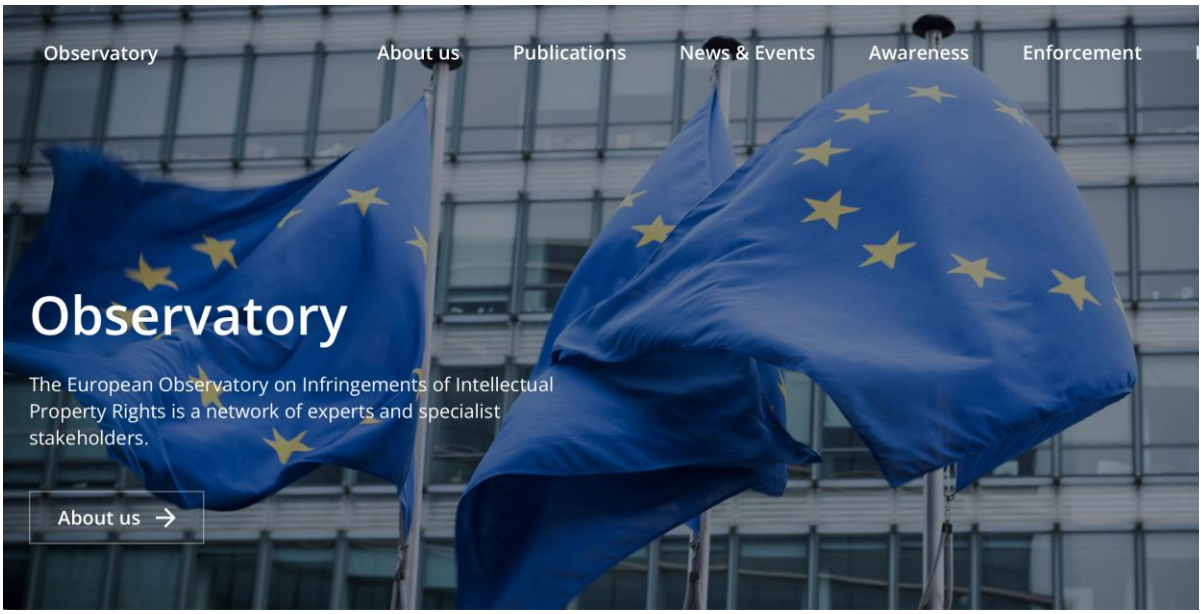


European Audiovisual Observatory

- Link: https://www.obs.coe.int/en/web/observatoire/
- Publications: regular, in PDF and xlsx document

2.4.3 European Observatory on Infringements of Intellectual Property Rights (EUIPO)

“The European Observatory on Infringements of Intellectual Property Rights is a network of experts and specialist stakeholders that brings together representatives from EU bodies, authorities in EU countries, businesses and civil society. The aim of the observatory is to improve the fight against counterfeiting and piracy by sharing information and best practice, raising public awareness, strengthening cooperation, and developing better tools.” [https://single-market-economy.ec.europa.eu/industry/strategy/intellectual-property/enforcement-intellectual-property-rights/european-observatory-infringements-intellectual-property-rights_en]



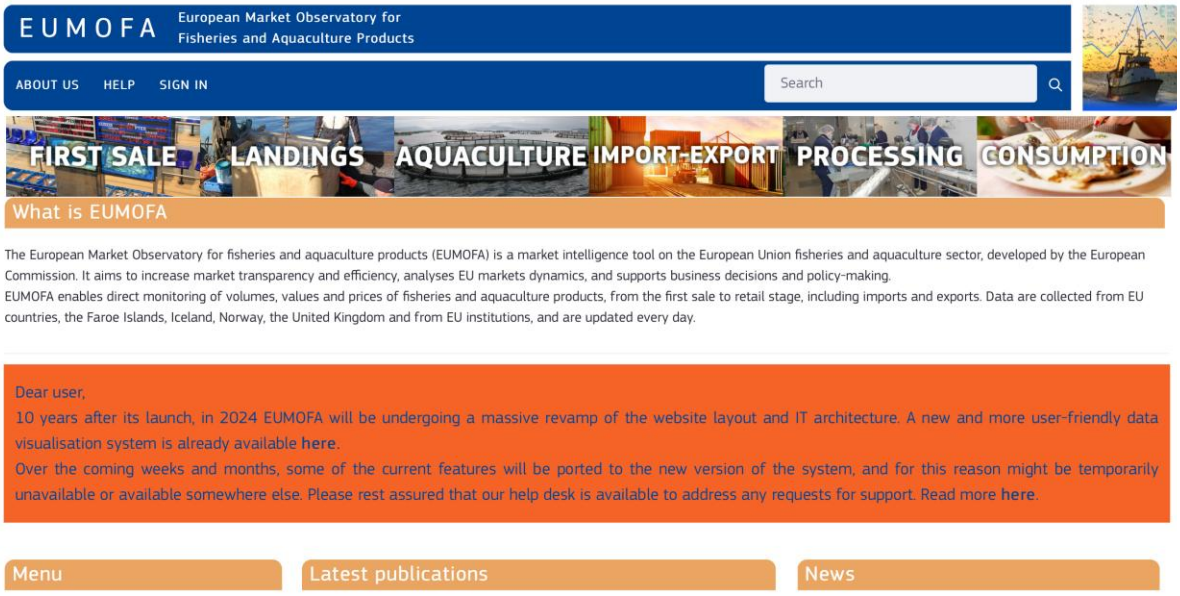
European Observatory on Infringements of Intellectual Property Rights

- Link: https://www.euipo.europa.eu/en
- Publications: regular, in PDF document
- Statistical data access: Several, under “Services” Menu
- EU Open Data portal: yes, via Directorate-General for Communications Networks, Content and Technology

2.4.4 European Market Observatory for Fisheries and Aquaculture Products (EUMOFA)

The European Market Observatory for fisheries and aquaculture products (EUMOFA) is a market intelligence tool on the European Union fisheries and aquaculture sector, developed by the European Commission. It aims to increase market transparency and efficiency, analyses EU markets dynamics, and supports business decisions and policy-making.

EUMOFA enables direct monitoring of volumes, values and prices of fisheries and aquaculture products, from the first sale to retail stage, including imports and exports. Data are collected from EU countries, the Faroe Islands, Iceland, Norway, the United Kingdom and from EU institutions, and are updated every day.



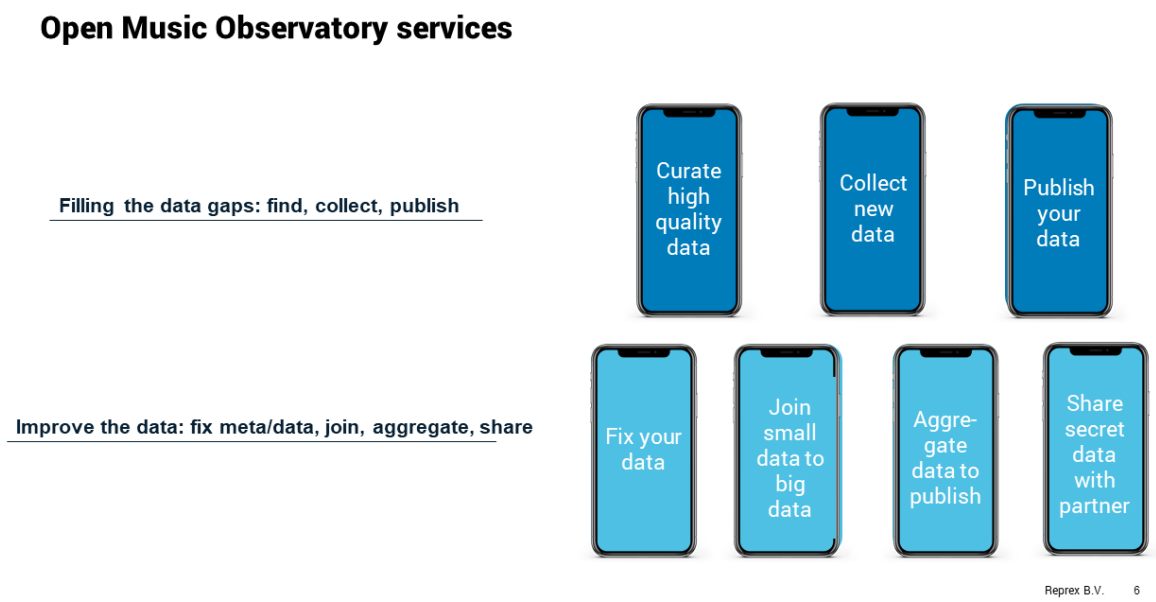
European Market Observatory for Fisheries and Aquaculture Products (EUMOFA)

- Link: https://eumofa.eu
- Publications: regular, in PDF documents.
- Statistical data access: EUMOFA DATA
- EU Open Data portal: yes, via Directorate-General for Maritime Affairs and Fisheries

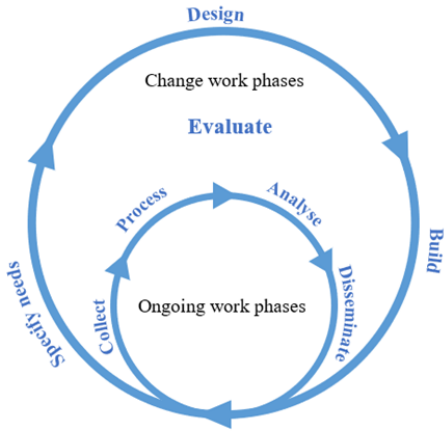
3 Core Services

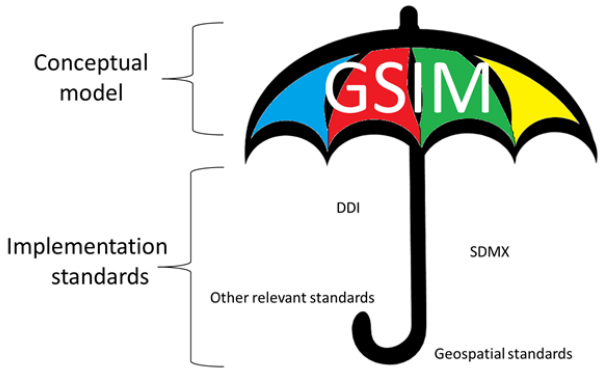
This section introduces the core services of the *Open Music Observatory*. These services are described in the *EMO Feasibility study*, but we gave them a modern and subjective interpretation that relies more on the novel innovations of data regulation and data science. Our services fully embrace the EIF and EOSC (related) interoperability frameworks and reproducible research techniques that allow for a more timely, less costly, and higher quality data ingestion and processing than manual workflows. Almost all workflows are supported by open-source components, some of which came as the background of the Open Music Europe project, were developed by other projects, or are being developed in different tasks of this project. The introduction of the separate software components is not a subject of this report.

Our main services are related to the collection, processing and dissemination of data.



The **Generic Statistical Business Process Model** (GSBPM) is an international standard model that "describes and defines the set of business processes needed to produce official statistics." (UNECE 2019) A conceptual reference or information model accompanies this business process model, the **General Statistical Information Model** (GSIM)(UNECE 2014). We use the conceptualisation of GSIM so that our results will be similar in quality to official statistics; of course, similar processes allow us to create products that combine well with official statistical products. The Open Music Observatory is not only collecting and disseminating statistically processed data, but also collections datasets, i.e., structured microdata. The GSBPM and GSIM cover both because statistical business processes rely on collection-like datasets, such as registers, codebooks, and metadata thesauri.





The two implementation standards, DDI and SDMX ensures that the Open Music Observatory can work together with statistical offices, or respectable social sciences data repositories.

In this report of Task 5.1, we concentrate on the inner cycle of the business processes. The outer cycle, specification, design, and build are carried out in other work packages, as is most of the analysis phase. We provide further services, too, which could relate in GSBPM to the 9th Evaluate header, which allows for quality improvements.

The two implementation standards, DDI and SDMX will ensure that the Open Music Observatory can work together with statistical offices, or respectable social sciences data repositories, because both our business processes and the way we organise data is following their standards. The *Data Documentation Initiatve* (DDI), will ensure that we will remain compatible with official statistical microdata and metadata services and other social sciences archives, like GESIS, the official data archive of all European Commission-mandated survey research dating back over 50 years (Vardigan, Heus, and Thomas 2008). The application of SDMX ensures that our microdata and statistically processed data will be interoperable with official statistics of the UN, OECD, Eurostat, and national statistical services (Stahl and Staab 2018).

Our data improvements, go beyond improvements of statistical quality and application of GSBPM; we aim to fix and improve music industry datasets for rights management or digital curation.

The **data enrichment and improvement** are innovative solutions that are not part of the services of an open data portal or an observatory. We aim to offer these value-added services to create new value and therefore motivation for music industry data owners to work with the observatory.

- **“Fix-the-data”** means improving the data quality by finding or imputing missing values or finding and replacing erroneous data entries. In terms of metadata, adding further machine-actionable information to already existing datasets can improve their usability.
- **“Data linking”**, data fusion, or data matching means correctly joining data from different datasets (data sources.) We ensure that data coming from sources can be meaningfully joined together; the variables have consistent meanings, the codebooks applied are harmonised; the timeframe or geographical frame is consistent.

- **Aggregation services**: we turn your music-related datasets into statistical products or data publications. We clean, validate, and structure it to a format that it can be placed on the EU Open Data Portal, Europeana, or Wikibase for integration with Wikidata/Wikipedia.
- **Confidential data sharing**: our data sharing space can be used for confidential data sharing and cross-pollination (for example, looking up missing ISWC/ISRC identifiers or misspelt names in each other's datasets) without making the data public.

These *planned* services will be discussed in Chapter 7.

3.1 Collect: Data Curation & Collection

Data curation is the organisation and integration of data collected from various sources. It involves annotation, publication and presentation of the data so that the value of the data is maintained over time, and the data remains available for reuse and preservation.

Data can only be understood with the broader concepts of information and knowledge, because data in itself is unprocessed, raw knowledge, that cannot be understood. The *EMO feasibility study* intuitively defines data gaps without an apparent reference to a data or conceptual model, but it recognises and stresses the need for terminological harmonisation.



Four types of data-collection principles have been identified as essential both by various branches of the music sector and also by policymakers at European, national and local levels:

- The data-collection service provided by a European Music Observatory should help in mapping, understanding and analysing the main characteristics, trends and idiosyncrasies of the music sector in Europe;
- The data collected should be neutral and available to decision-makers, music sector operators, and the public;
- The data itself should cover the activities of the music sector across the entire European Union, be comparable between Member States, and rely on identified and stable indicators;
- The data collection methods should be transparent and provide a strong degree of scientific accuracy. (European Commission et al. 2020, p28)

In short, we collect data about music, as defined in the cultural statistics of any European Economic Area and EU candidate statistical office or by a representative European or international music organisation.

In more detail, we a systematic data collection program requires a *conceptualisation* is an abstract, simplified view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them.



Usually, when we record information about a musical work, we do not make a copy of the entire work but record some identifying properties of the work, for example, the name of its author and the name (i.e., the title), its unique ISWC identifier, and the data or registration. We work with a *concept* of a musical work, not with the entire work.

Composers as human beings are represented by their names, IP Names or ISNI identifiers, and date of birth and death. Again, in an information system we obviously work with a *concept* of an author, and *instances* of authors represented by their unique data.

The *EMO feasibility study* catalogues 45 data gaps that a future European music observatory should fill. A data gap can only be formally defined and filled with some reference to conceptual models of the world. A typical data problem plaguing the music sector is the amount of computer and human work needed to connect musical works and their recorded fixation, and eventually, the composers, producers, and performers linked to these objects for royalty payment. We need agreed concepts of the *composer*, *sound recording*, *work*, to answer such questions.

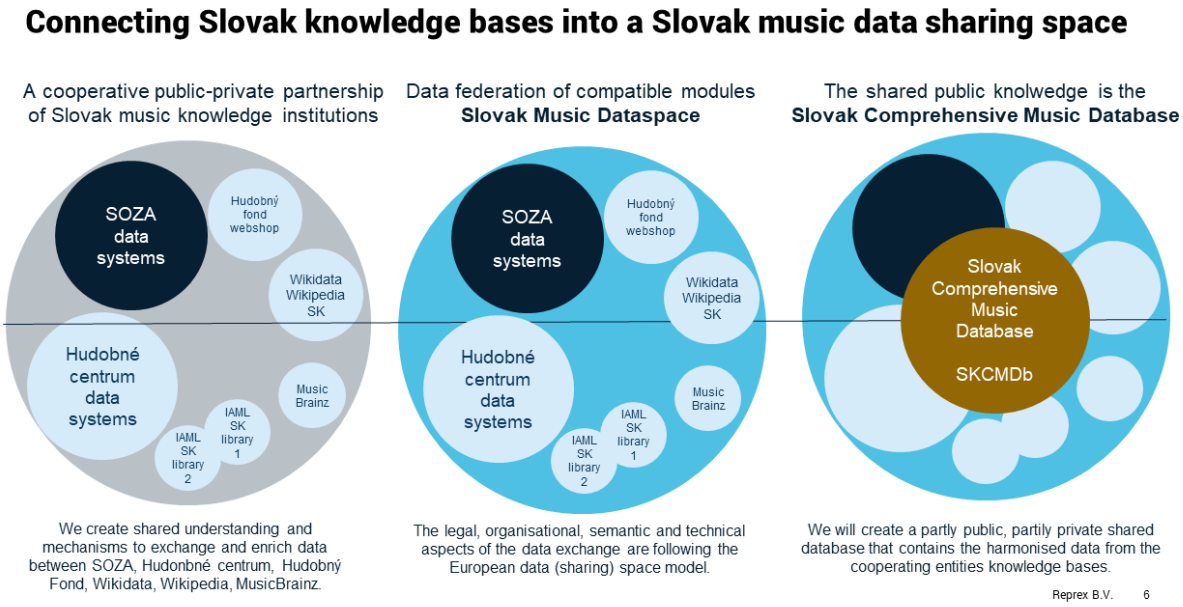
The initial data collection guidelines of the Open Music Observatory are derived from the EMO Feasibility study. We see them as a starting point for further discussion with the Observatory Stakeholder Network. We introduce them with our data catalogue in Section 4.1. These guidelines are supported by our first conceptualisation, which is built on some widely used conceptualisations of creative works and statistics. This is the topic of Chapter 6.

3.1.1 Microdata, Collections, Records

We treat “microdata” as a collection of structured data. A **collection** is a group of objects, for example, musical works, sound recordings, music enterprises, musician biographies, gathered together for some intellectual, artistic, or curatorial purpose. This is how radio playlists and charts, festival line-ups, local content guideline monitoring works. Such collections form the basis of census or sample surveys for statistical data collection.



Our first large collection is the **Slovak Comprehensive Music Database**. Our aim is to publish a constantly refreshed database of every music composed or recorded in the territory of the current Slovak Republic, or composed and recorded by people from Slovakia, or sung in the Slovak language. This database will be based on a collection of musical works and a connected collection of sound recordings and sheets as manifestations of those works.



- ☒ Our collections are always available on openmusic.wikibase.cloud↗. Further details in Section 3.3.3.
- ☒ Whenever our collections fit in the collection and publication guidelines of Europeana, we make the collections available there, too.
- ☐ We are investigating the possibility of synchronising our collections to the European Cultural Heritage Cloud.

The *Data Documentation Initiative* is originating for the world of social sciences data archives and more and more in use in statistical organisations for the documentation of microdata. The DDI plays a particularly important role in the creation of statistical surveys, particularly using questionnaires and question banks. The new *Records in Context* has replaced the international standards on archives in 2023. Its central concept is the *record*, which is a document according to DDI; a collection is a *set of records*. Our standardisation of microdata is explained in more detail in Chapter 6.

3.1.2 Primary data collection

The Open Music Observatory is supporting high-quality primary data collection, and itself is carrying out such collection activities.

The indicators derived from the processing of survey questionnaires will be comparable if the same concepts of interest (for example, concert visiting frequencies) are measured via the same questions and answering instructions.

> ⓘ
>
> A concert is a standard concept of a live performance of music.
>
> How many times in the previous [12 months] have you been to a concert? is a standard question accompanied by standardised answer options and processing in the Cultural Access and Participation surveys following the ICET model.

Using standardised concepts and question banks, including question and instruction labels with standardised translations, is a cornerstone of ex-ante survey harmonisation. This process is a prerequisite for retrospective survey harmonisation and the subsequent creation of comparable statistical indicators, underscoring the importance of uniformity in data collection.

- ☒ We provide API and download access to harmonised, multi-language question banks. This allows music stakeholders to use the same question formulations and translations for comparability with European statistical and policy research programs.
- ☒ We provide tutorials to *retroharmonize*, a background open-source software of Reprex, which is an R library to retrospectively harmonise data from different survey programs that had asked the same questions.
- ☒ The Open Music Europe project will carry out some harmonised surveys to show and improve the methodology of harmonised data collection within the music sector of Europe. This data will be available as metadata (questionbank), as microdata (individual answers), and as processed statistical data.

3.1.3 Metadata

The most common—and perhaps least useful—definition of metadata is that it is "data about data." As catchy as this definition is, however, it is entirely ambiguous. First of all, what is data? And second, what does "about" mean? (Pomerantz 2015, p19)

The new ISO standard on Information technology — Metadata registries (MDR) defines *metadata* as data that define and describe other data. As Pomerantz eloquently argues, this is a definition that is not very helpful. We use his more functional (but not contradictory) definition. "Data is only potential information, raw and unprocessed, prior to anyone actually being informed by it. […] Data must be understood not as an abstract concept but as objects that are potentially informative. […] Metadata Is a Statement about a Potentially Informative Object." (Pomerantz 2015, p26)

Following the metadata definition of "a statement about a potentially informative object," we believe that any high-quality data can be used as metadata in certain circumstances.



Data or metadata?

The data of birth can be seen as a metadata for disambiguation among authors with the exact same name in a copyright register. It can be seen as data for a curator of a young author prize, or a music sociologist. Either way, the date of birth should be precise, and encoded in a way that makes it portable and interoperable.

From a data management point of view, we do not distinguish between data and metadata. Of course, we acknowledge the fact that some types of data will always remain under the hood and will only serve the proper functioning of an information system.

The music industry's famous "metadata problems" usually arise when a music enterprise or institution wants to use metadata information from an authoritative source that is somehow corrupted. The Open Music Observatory can help with these metadata problems by disseminating proper, open authoritative data (as registers or collection) or by providing data improvement services that fix the metadata problems of a user.

3.1.4 Statistical indicators and datasets

⚠

We will place our first statistical datasets to the EU Open Data Portal this week (pending their approvals) and will provide a screenshot and access conditions here.

3.2 Process

We use the theory of metadata by Jeffrey Pomerantz, who defines Metadata as “a statement about a potentially informative object.” A dataset without such statements is not findable, accessible, interoperable, and very hard to reuse. Pomerantz distinguishes among descriptive, administrative, structural, preservation, and use metadata. The Generic Statistical Information Model (GSIM) is a common abstract representation of data objects manipulated in official statistical production and elaborated as an overarching model for implementation metadata standards such as SDMX or DDI.

GSIM since its inception aims to bridge two important standards, SDMX and DDI. The Statistical Data and Metadata Exchange has been developed for decades and it is an ISO standard; it is more geared towards the aims of data sharing and preservation in RDM. DDI on the other hand is more focused on the documentation and quality control of primary data collection, or the reuse of often messy data sources, and supports the processes that make the data available for research. As DDI provides information about a much wider range of objects and processes, we are even more selective when we turn to this standard than SDMX; however, we cannot disregard DDI for microdata.

3.2.1 Processing & re-processing microdata

⚠

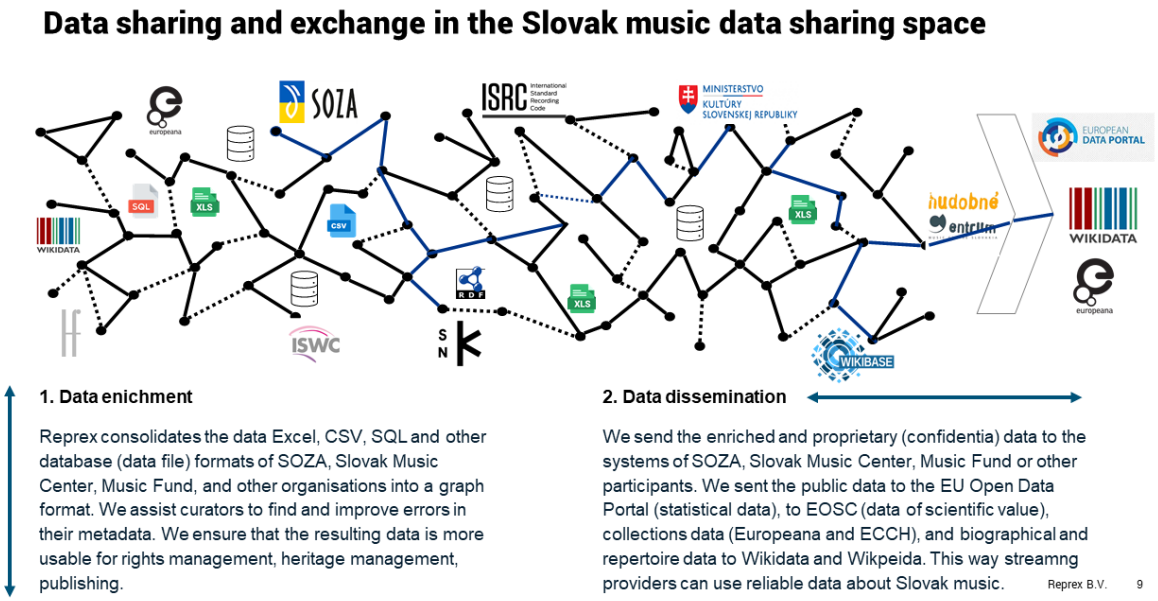
We will place here an example that goes to the EU Open Data portal

3.2.2 Documentation

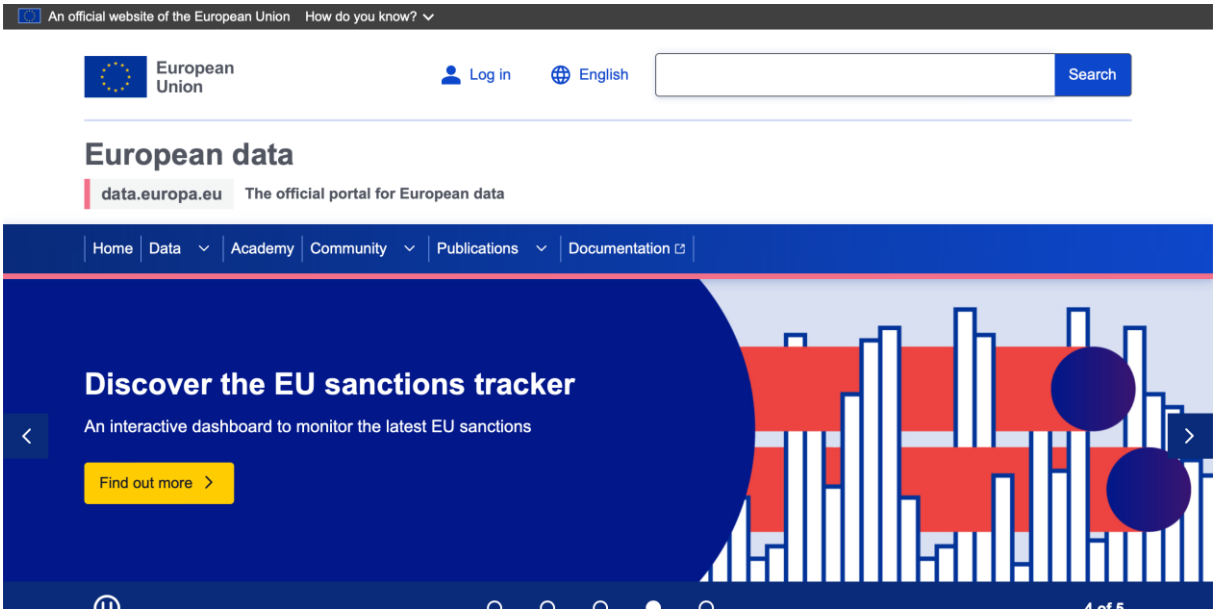
⚠

We will provide the link and screenshot of the documentation for each file that goes public.

3.3 Disseminate



3.3.1 EU Open Data Portal



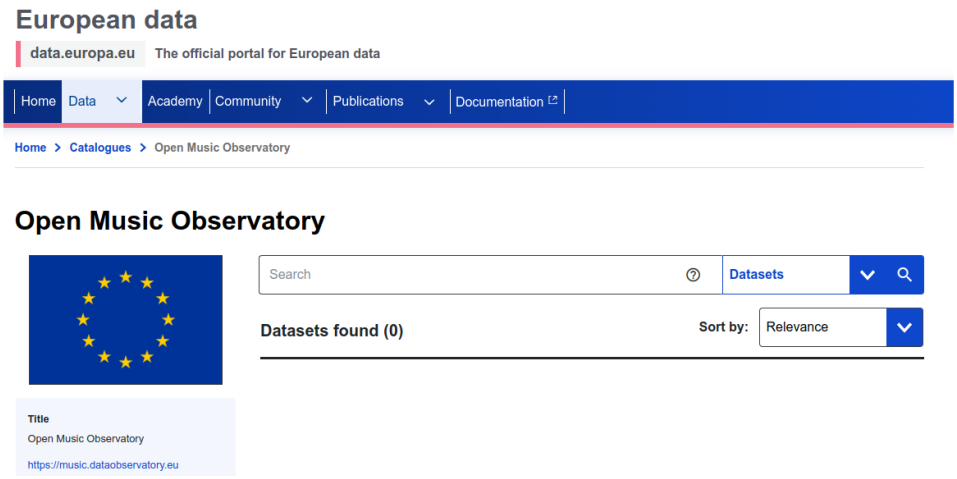
The portal is intended to:

1. give access and foster the reuse of European open data among citizens, business and organisations.

2. promote and support the release of more and better-quality metadata and data by the EU's institutions, businesses, agencies and other bodies, and European countries, enhancing the transparency of European administrations.

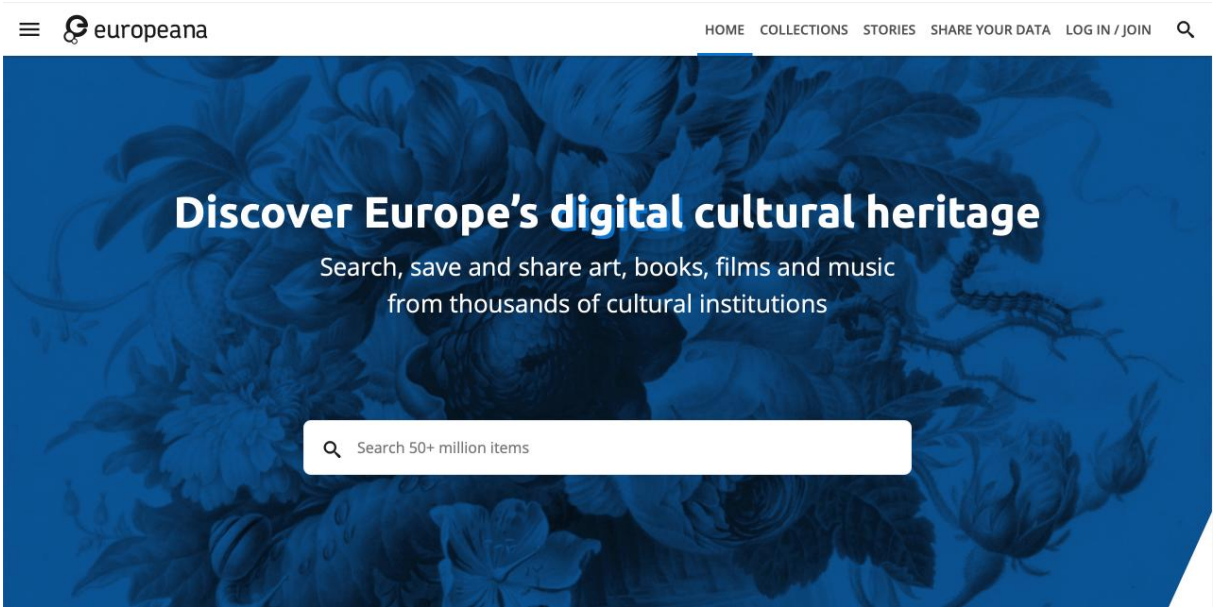
3. educate citizens and organisations about the opportunities that arise from the availability of open data.

It is funded by the EU and managed operationally by the Publications Office of the European Union in cooperation with the *Directorate-General for Communications Networks, Content and Technology* of the European Commission, responsible for EU open data policy.



We publish our data primarily on the EU open data portal for statistically processed datasets (datasets that contain the generalised characteristics of many data subjects without personal data that could identify them).

3.3.2 Europeana



Europeana is a bottom-up, decentralised data aggregator for the cultural heritage part of the music sector and the broader cultural and knowledge sector.

Europeana is at the heart of the common European data space for cultural heritage, a flagship initiative of the European Union to support the digital transformation of the cultural heritage sector. Millions of cultural heritage items from over 3,500 data providers across Europe are available online via the Europeana website. We work to share and promote this heritage so that it can be used and enjoyed by educators and researchers, creatives and culture lovers across the world.

While there is no agreed, cross-sectoral definition of “collections”, it is widely understood that in many cases, collections themselves are the entities that meet the information needs of music professionals or researchers (Wickett et al. 2013). The creation of collections is an important activity performed by music professionals and scholars as part of their work process. For example, if we want to measure how many European or French works made it ever to an American hitlist, we have to contrast two collections (French works, and the collection of works that were ever on the particular hitlist) to calculate this indicator.

The publishing policies of Europeana are very restrictive, and therefore, there currently needs to be more musical works on this important open knowledge graph. After consultation with Europeana Sound, the British Library-based aggregator responsible for the music in the European collection, we decided to pursue two ways to make more extensive European music collections visible.

We will publish collections with publicly viewable audiovisual material on Europeana via the Open Music Observatory’s collections. We will also start a discussion with Europeana’s new project, the European Cultural Heritage Cloud, which has less restrictive data licensing policies, about a more extensive dissemination point for our collection datasets.

3.3.3 Wikibase Cloud

Our main dissemination point for microdata is openmusic.wikibase.cloud, available both from our website and via the main Wikibase Cloud website. Wikibase Cloud is an initiative of Wikipedia and Wikidata to bring more specialised data into the Wikimedia ecosystem.

Wikibase is a software system that help the collaborative management of knowledge in a central repository. It was originally developed for the management of Wikidata, but it is available now for the creation of private, or public-private partnership knowledge graphs. It was developed by Wikimedia Deutschland.

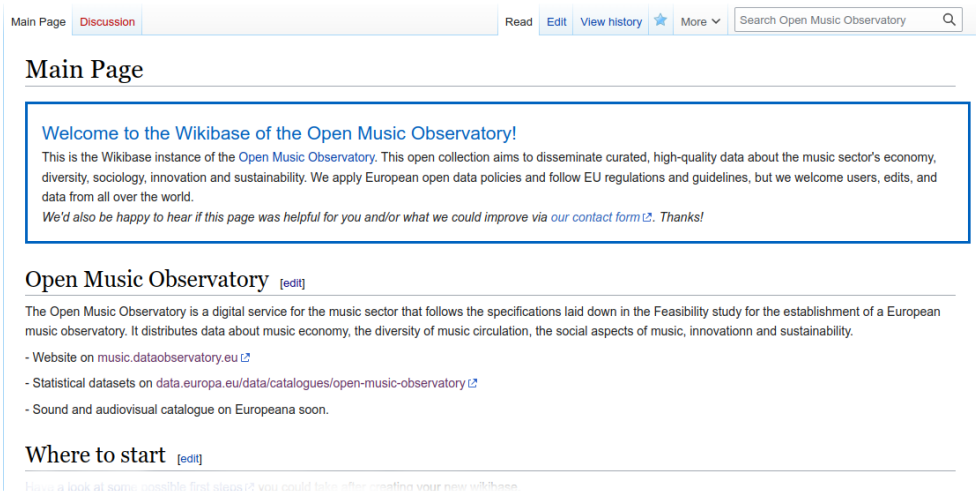
Wikidata itself is a gigantic *Wikibase instance*. Their user interface is similar, but depending on what the administrator of your Wikibase instance allows you to do, you are likely to have more freedom to edit certain elements, like properties, than on Wikidata. Wikidata must protect the integrity of one of the world’s largest knowledge systems, and does not allow editing access to certain elements.

Because of the success of Wikidata, several EU projects and institutions started to use Wikibase, the software that runs Wikidata. They aim to reuse the software to construct institutional or cross-institutional, domain-specific knowledge graphs. Several factors make Wikibase attractive:

- ☒ the fact that it is a well-maintained open-source software;
- ☒ there is a rich ecosystem of users and tools around it;
- ☒ Wikimedia Deutschland↗ (WMDE), the maintainer of Wikibase, has made considerable investments in optimising the software’s use outside of Wikidata or other Wikimedia projects;
- ☒ The EU Knowledge Graph↗ runs on Wikibase;

- ☒ The *EU Academy* and the *EU Open Data Portal* actively disseminate good practices and know-how on its implementation in cross-institutional data-sharing programs.

Our main dissemination point for non-statistical data is the Wikibase Cloud.



3.3.4 Music Observatory Website

As of this deliverable's resubmission, the OMO is available at https://openmusicobservatory.eu/



3.3.5 API Endpoint

The OMO API endpoint will be described in the “Instructions” section of the website: http://135.181.91.51:3008/en/instructions

4 Data Catalogue

The Open Music Observatory curates, maintains, and disseminates a data catalogue with the resources within the data catalogue: individual datasets and their series and API endpoints where the data can be queried in a custom format. In creating our data infrastructure, we considered the specifications of our dissemination nodes, which provide our data with a wide range of interoperability and easy access: the EU Open Data Portal, Europeana, Wikibase Cloud and Wikidata. From a thematic point of view, we relied on the definition of the *EMO feasibility study*, and created topical pillars (Section 4.2). The data curators of the Observatory Stakeholder Network (see Section 8.1.1) and for the duration of the Open Music Europe project, the work packages (WP1-4 represent each "pillar") can define and provide datasets or data series according to their topical collection guidelines (Section 4.1).

A data catalogue formally is a metadata dataset: a dataset on information about our available datasets and their downloadable or queriable distributions. It follows the global World Wide Web DCAT standard.

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. This document defines the schema and provides examples for its use (Albertoni et al. 2020). It is a global standard, which was further extended and specified for the release of statistical datasets (StatDCAT-AP) and for the needs of the EU Open Data Portal (DCAT-AP). (Sofou and Dragan 2019; Fragkou 2023) These extensions provide further metadata and organisations standards, but essentially they do not change the definition of the global standards.

A data catalogue (dcat:Catalog) represents a catalogue, which is itself a dataset in which each individual item is a metadata record describing some resource: a description of a dataset, a data service, or other type of resource.

dcat:Dataset represents a collection of data, published or curated by a single agent or identifiable community. We currently support two types of datasets: statistical datasets that conform to the datacube definition of SDMX, or collection datasets for microdata, which contain non-aggregated, structured data representing some unity criteria, for example, music works and recordings that have been present in the official radio charts of a given country. The _dataset_, similar to a musical or literary work, is an abstract concept which can be used, downloaded, and stored in its manifestation. For a musical work, a manifestation may be a sound recording or music sheet; for a dataset, it is a distribution. A URI identifies a dataset; the URI does not allow the downloading of the dataset, because it refers to the abstract idea of the dataset; the URL for downloading the dataset belongs to the individual distributions.

dcat:Distribution represents an accessible form of a dataset, such as a downloadable file. When the same dataset is distributed in different file formats (for example, CSV and SPSS files), each distribution is listed in the catalogue separately with a separate download link. Each distribution has its own URL where the dataset can be downloaded.



The music.dataobservatory.eu/tag/music-economy/ URL lists the downloadable datasets on the Open Music Observatory website. They can be found on EU Open Data Portal, too.

In the first days after launching our new service, around 1-10 June 2024, the datasets may be missing from the EU Open Data Portal, which is changing in these days its complete backend, and may have some backlog in accepting our datasets.

dcat:DataService represents a collection of operations accessible through an interface (API) that provides access to one or more datasets or data processing functions. Our datasets are accessible on different platforms with their own datasets, and our internal data-sharing space also has its API. As data is added to the different platforms (EU Open Data Portal for statistical and microdata datasets, Europeana for collections dataset, Wikibase Cloud for further microdata, metadata and collections, and Reprexbase for confidential microdata and collections), we are updating the catalogue with the DataService entries.

dcat:DatasetSeries is a dataset that represents a collection of datasets that are published separately but share some characteristics that group them; for example, a (play)list of sound recordings that were present in the weekly charts or the annual budget of an institution. A time series dataset is usually not defined as a data series, but the new time observations are added to an updated distribution of the time series dataset. Stakeholders who provide data to the Open Music Observatory can commit to making a data series; however, we only define a data series when we have at least two items available from the series.

dcat:CatalogRecord represents a metadata record in the catalogue, primarily concerning the registration information, such as who added the record and when.

4.1 Collection Guidelines

In short, we collect data about music. The initial data collection guidelines of the Open Music Observatory are derived from the EMO Feasibility study. We see them as a starting point for further discussion with the Observatory Stakeholder Network.

- ☒ Statistical data which is defined as cultural statistics of any European Economic Area and EU candidate statistical office or by a representative European or international music organisation.
- ☒ Statistical data (indicators and their datasets) defined by, or requested by members of the Observatory Stakeholder Network.
- ☒ Datasets about information gaps identified by the *EMO feasibility study*.
- ☒ Records of questionnaires, question banks, and any structured datasets used for the creation of the statistical datasets above.
- ☒ Collection datasets about musical works and their manifestations in sound recordings or musical sheets.
- ☒ Collection datasets about music events, including events of composition, recording, or live performance.

- ☒ Encyclopaedic, demographic, biographical data about music professionals and music enterprises.
- ☒ Collection datasets about books, publications, statutes and laws, standards related to music.

The *EMO feasibility study*

Curators are forming collections with the application of unity criteria which allow them to decide which musical work, sound recording, music enterprise or person is included in a collection list. The curators are responsibility for the comprehensive application of the unity criteria and ensuring that their collections are up-to-date (Wickett et al. 2013).

Some examples of music data curation

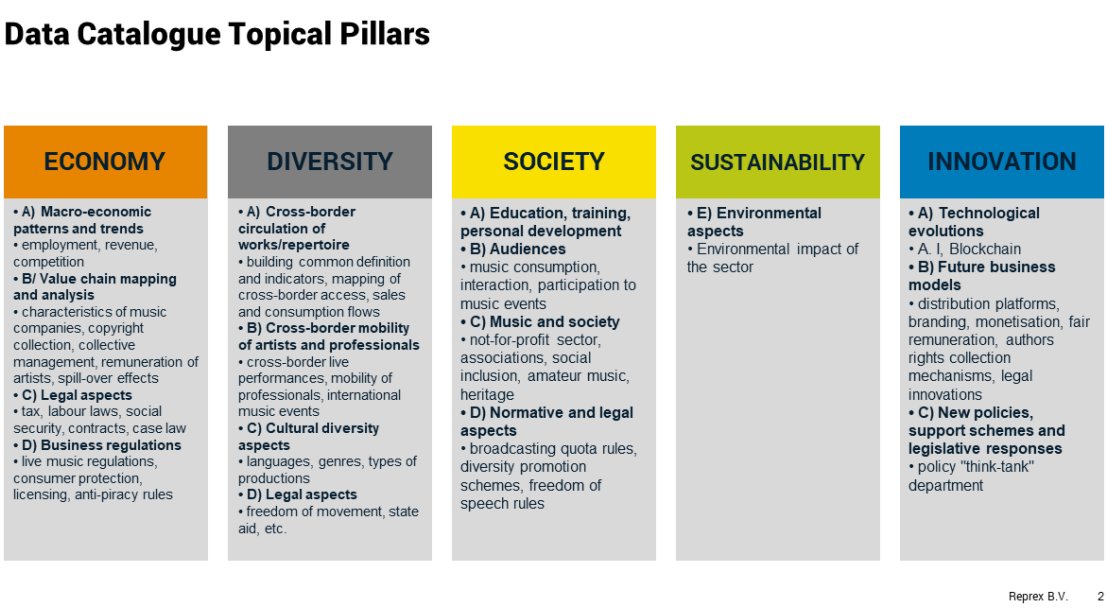
Hitlists use some kind of popularity metrics, and they follow rigorous rules which sound recordings are included every week, or year.

Collective rights management organisations create comprehensive lists of works and sound recordings registered for rights protection and exploitation.

Statistical agencies create business registers to carry out data collection.

Who can curate our datasets? Any music professional or scholar can curate datasets in agreement with our Collection Guidelines. The quality review mechanisms will be set by the Observatory Stakeholder Network from a content point of view, and the Open Music Data Exchange from a technical point of view.

4.2 Topical Pillars



Sustainability requires partially different data sources compared to other social and economic indicators (with the exception of certain ESRS European Sustainability Reporting Standards, social

dimension S indicators). Therefore, the data catalogue distinguishes five topical areas, including sustainability, while maintaining a four-pillar presentation structure consistent with the Feasibility Study.

The suggested four-pillar model would categorise data-collection and analysis along the following lines:

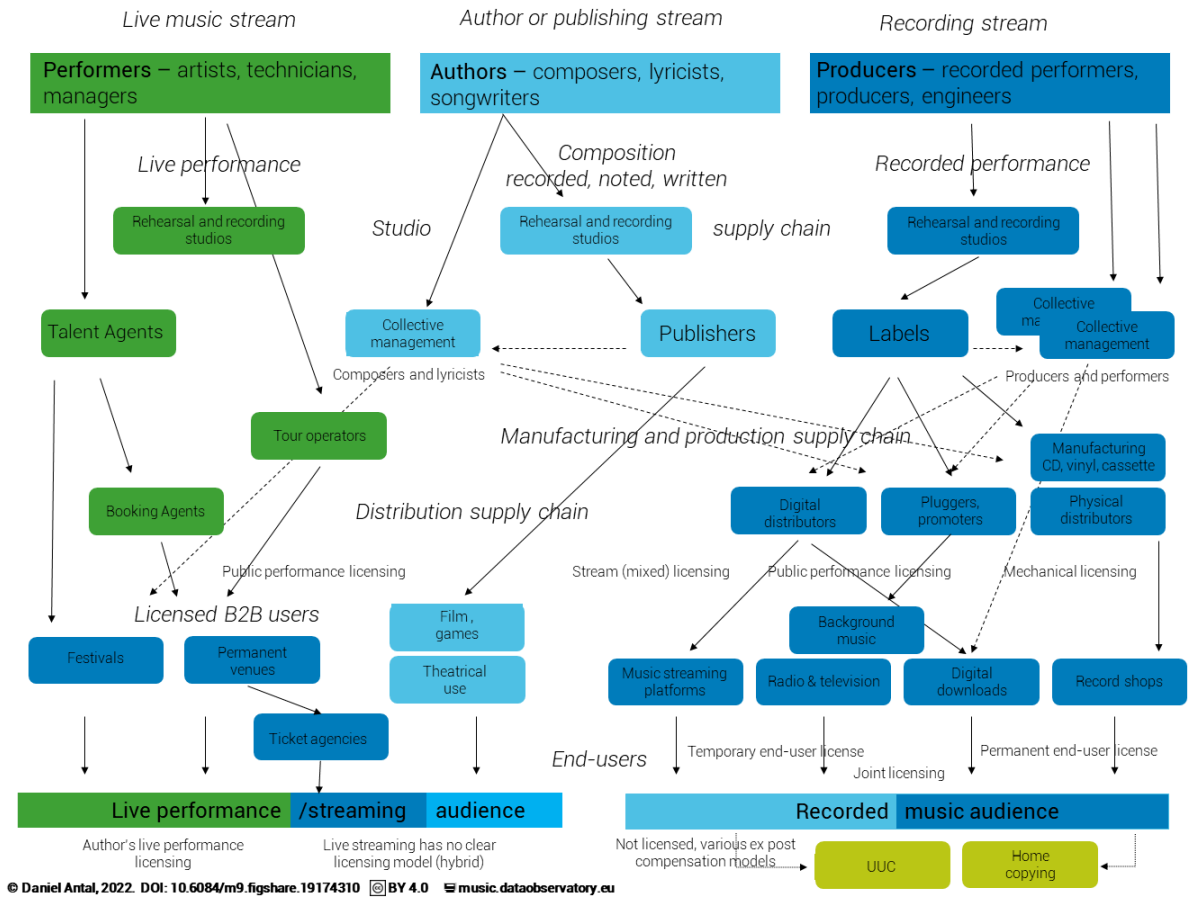
- Measure the contribution of music to the EU's economic and legal environment, from a systemic perspective (Pillar 1).
- Monitor the cross-border flows of repertoire, the mobility of artists and diversity (national, linguistic, genre-based) (Pillar 2).
- Assess music's impact on society and citizenship: how audiences access and consume music; how citizens participate in professional and not-for-profit music activities; the scale, value and quality of music education and training (Pillar 3).
- Provide a framework to develop prospective research on the future of the music sector, supporting innovation and developing understanding of emerging practices from various perspectives (business, tech, policy) (Pillar 4)(European Commission et al. 2020, p30).

4.2.1 Music Economy

Main potential data-collection and research areas identified at this stage: ▷ Macro-economic patterns and trends (e.g. employment, revenue, competition) ▷ Value chain mapping and analysis (e.g. characteristics of music organisations, copyright collection, collective management, remuneration of artists, spill-over effects) ▷ Legal aspects (e.g. tax, labour laws, social security, contracts, case law) ▷ Business regulations (e.g. live music regulations, consumer protection, licensing, anti-piracy rules) (European Commission et al. 2020, p114)

According to the *EMO feasibility study*, one "of the key findings of the AB music working group report was a substantial appetite for cross-sectoral, neutral and comparable data on the music business at EU level. While recent studies (e.g. EY "Creating Growth" study) have attempted to measure the impact of music on the EU's economy, systematic and comprehensive metrics do not exist at this stage." (European Commission et al. 2020, p113)

Deliverable D1.1 of Open Music Europe, *Economy of Music in Europe: Methods and Indicators* identifies critical research questions, data sources and gaps, and data collection methods regarding the economy of music in Europe . (Antal, Kmety Barteková, and Remeňová 2023) The deliverable begins by reviewing definitions of "the music industry", the categorisation of musical activities within the system of national accounts (SNA) and statistical classifications of economic activity (ISIC and NACE), and the three primary income streams within the music industry (the live music, author or publishing, and recording streams).



It then turns to the topic of value, first identifying the types of value created by musical activity and then considering legal and economic dimensions of valuation.



Our website documents with visualisations each dataset, apart from providing links to the latest distribution downloads with visualisations (in zip) or the access points on the various dissemination nodes. The illustration is an experimental dataset from our background CEEMID catalogue.

After introducing the concept of mixed enterprise and personal surveying as a means of improving insight on informal economic activity in the sector, the deliverable identifies data gaps relevant to national policy in our pilot study target country of Slovakia, critically reviews the data gaps relevant to EU-level policy first identified in the *EMO feasibility study*,and proposes data collection methods appropriate to filling specified data gaps.

4.2.2 Music Diversity

According to the *EMO feasibility study*, "creating reliable tools to monitor what kind of repertoire circulates on digital platforms or via radio will require access to vast amounts of data from Digital Service providers (DSPs) or third party aggregators. The notion of European repertoire has to be clarified and very well defined; notion of language, of origin, of nationality, country of production, genres, and it should not be limited to the language sung in each song. […] A European Music Observatory should also look into the possibility to collect regular data on the circulation of European repertoire at song and/or artist level, considering live performance/radio/ digital use, which will be available at a weekly/monthly/yearly basis to the music sector." (European Commission et al. 2020, p34)

The Open Music Europe project is developing two tools for capturing and turning the aforementioned data into informative indicators. In WP1, the project is developing big data statistical sampling algorithms to avoid the need for "access to vast amounts of data from Digital Service providers (DSPs)". WP2 is working on a taxonomy and GDPR-conform representation of the "notion of language, of origin, of nationality, country of production, genres" based on our background (Antal 2020b). The result of this work will be the Slovak Comprehensive Music Database, which will create clear taxonomies and allow users and software applications to determine aspects of "Slovakness" for each sound recording.

The possibility "to collect regular data on the circulation of European repertoire at song and/or artist level" is an issue that must be addressed with strict adherence to GDPR. We are developing an opt-in, opt-out mechanism in Slovakia that will be replicable in all other jurisdictions in accordance with GDPR. (For potential non-European artists, we will apply GDPR, too.)

> Main data-collection and research areas identified at this stage: ▷ Cross-border circulation of works/repertoire (e.g. building common definition and indicators, mapping of cross-border access, sales and consumption flows ▷ Cross-border mobility of artists and professionals (e.g. cross-border live performances, mobility of professionals, international music events) ▷ Cultural diversity aspects (e.g. languages, genres, types of productions) ▷ Legal aspects (freedom of movement, state aid, etc.) (European Commission et al. 2020, p115)

An interesting opportunity here is that many European countries already collect information on subsidised music operators (e.g. associations or not-for-profit projects), not to mention the wealth of information available through Creative Europe supported initiatives, which could provide this Pillar with interesting data. (European Commission et al. 2020, p116)

4.2.3 Music Society & Sustainability

Regarding music and society, D3.1 considers the reuse of various survey programs with retrospective survey harmonisation, and WP3 is planning to conduct surveys in 2025. Data will be added to the catalogue as it is becoming available.

> ⓘ
>
> Main data-collection and research areas identified at this stage:
>
> - Education, training, personal development
> - Audiences (music consumption, interaction, participation in music events, etc.)
> - Music and society (not-for-profit sector, associations, social inclusion, amateur music, heritage, participation in music)
> - Normative Aspects (broadcasting quota rules, diversity promotion schemes, freedom of speech rules)
> - Music and the environment (carbon footprint of venues, touring, festivals, merchandise manufacture, streaming services; issues around noise/neighbourhood impacts; good practice in these areas). (European Commission et al. 2020, p116)

In the *EMO feasibility study* the definition of sustainability was mentioned among the innovation topics. Because of the triple transition, introducing the Corporate Social Responsibility Directive and the European Sustainability Reporting Standards have increased the interest and need in sustainability data; we decided to create a separate topical area for environmental and social sustainability, or governance indicators (ESG.) We will publish datasets that will be used in the value-added service described in Section 7.2.2.

4.2.4 Innovation

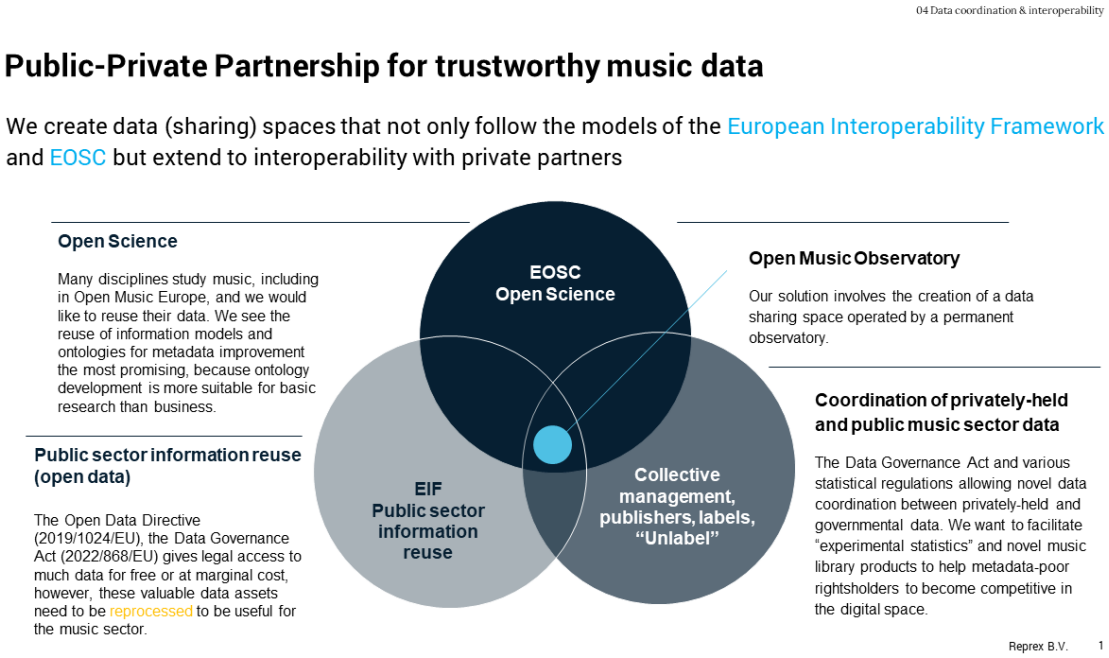
The definition of the Innovation pillar in the *EMO feasibility study* is more a topic to be covered than a data need description.

This pillar is less data-driven in that it will rely mostly on research conducted on topics relating to changes in the marketplace, new business models, disruptive technologies, etc. A European Music Observatory will have the latitude to pick certain topics based on priorities and input from sectoral stakeholders. An EMO should consider setting up an "innovation experts' advisory committee," constituted of respected professionals in their field who are known for their forward-thinking views, to help identify key themes to be studied. (European Commission et al. 2020, p37)

We will initiate an informal music innovation expert's roundtable to discuss potential data needs in this pillar.

© 2024 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295

5 Data Sources



In general terms, three main situations have arisen in the analysis of availability of data covering the music sector:

1. Data is available through stakeholders and would be supplied to the EMO at no cost or, if needed, at the cost of processing the data. The real cost would then be that of the human resources necessary to analyse and present the data.

2. Data is available through vendors whose business model is to sell or license data, research and analysis. The data would be made available to the EMO following a commercial and contractual negotiation, the terms of which were not available for this report.

3. Data is not available or not tailored for the needs of the EMO and therefore, access to such data would require EMO to establish the conditions for this data to exist, in partnership with stakeholders and data suppliers, at a cost that is difficult to determine without evaluating exactly the task at hand and the cost-benefits of developing such data. (European Commission et al. 2020, p62–63)

5.1 Curation of Reusable Data

Data is available through stakeholders and would be supplied to the EMO at no cost or, if needed, at the cost of processing the data. The real cost would then be that of the human resources necessary to analyse and present the data.

The *EMO feasibility study* did not discuss data curation; in our view, it has a rosy picture of data reuse. We believe that the Open Music Observatory should only ever ingest data with reprocessing; even

when pristine data arrives from an official statistical source, it needs proper data provenance documentation to synchronise with potential later revisions or corrections by the original authority.

“Data available through stakeholders”, in the case of the many music stakeholders mentioned in the EMO feasibility study, almost always requires significant curatorial investment. In our experience with around 100 music stakeholders, only a few, for example, government-sponsored, larger music information centres or developed music market collective management agencies, employ trained staff for data curation.

Our first use case in Slovakia was built around the Music Center Slovakia because it is an organisation with a competent in-house library and an IT team. Agreeing on a data curatorial workflow with trained librarians is easy if their data is well maintained. In the music sector, where businesses and social enterprises usually have less than five staff, no trained library (information) scientists or in-house IT technicians are present.

Our data curation workflows, therefore, are aligned for three distinct types of stakeholders:

1. Data curation excellence centres like the EU Open Data Portal team, Europeana, Wikibase, Eurostat, national libraries, and official statistical offices have competencies in linked open data and conceptual and information models, and flawless data coordination is expected.
2. Data expert organisations, like collective rights management agencies, digital distributors, larger music publishers and labels, and music information centres, who have a competent IT team and experience with reconciling their local databases with external databases.
3. Industry specialists and music scholars who do not have in-house or specialised IT capacities often work in spreadsheet (Excel) databases. They need training or full data curation support.

Our technological choices are explained in Section 3.3.3, and in further detail in @. Our preference for building our system around Wikibase, the software that coordinates the 329 language version contents of Wikipedia, as well as open knowledge graphs like Wikidata, Wikispecies, Wikiquotes, and other very extensive, linked data resources. Wikibase was also used successfully by the success of several EU and member state projects where Wikibase was the right choice for stakeholders who did not have the human/IT resources of data curation excellence centres. Wikibase is being developed with citizen scientists and domain subject experts with little IT background.

In the past years, Wikibase has proven to be an excellent bridging tool between data excellence centres, applying more sophisticated conceptual and information models and subject domain experts working with simpler systems. One aspect of this success is the Wikidata Data Model, which is based on the Resource Description Standard, like most international data curation standards, but offers relatively simple relationships among the models. This means that library or archive collection management systems require interfaces to import their data into Wikibase or export the data curated with music industry stakeholders to more complex models. We discuss these procedures among our data improvement services in Chapter 7.

5.1.1 Curation from data vendors

> Data is available through vendors whose business model is to sell or license data, research and analysis. The data would be made available to the EMO following a commercial and contractual negotiation, the terms of which were not available for this report. (European Commission et al. 2020, p39)

At this point we have not considered purchases from data vendors.

5.1.2 Novel data assets

> Data is not available or not tailored for the needs of the EMO and therefore, access to such data would require EMO to establish the conditions for this data to exist, in partnership with stakeholders and data suppliers, at a cost that is difficult to determine without evaluating exactly the task at hand and the cost-benefits of developing such data. (European Commission et al. 2020, p39)

All Work Packages of the Open Music Europe project will create novel methodologies and indicators; these experimental indicators will be added till 2025 December 31 to the data catalogue.

5.2 Data Providers

Organisations that indicated willingness to exchange data with a future European music observatory

Stakeholder	**Description**	Access for observatory
AEPO-ARTIS	Organisation representing European artists-performers. Regroups most of the European CMO representing performers.	Subject to non- financial partnership agreement. Would also require approval from SCAPR board.
CEEMID	Data collection and integration system based on open data, open- sources and online surveys.	Interested in contributing and working in partnership with a future EMO.
CISAC	Trade organisation regrouping rights societies in the world.	Subject to non- financial partnership agreement.
CNM (former CNV)	Public organisation managing a tax on concert tickets	Subject to non- financial partnership agreement.
DDEX	Standards-setting organisation regrouping all stakeholders in the digital value chain.	Interested stakeholder in particular contributing to the Innovation & New Models pillar.
GESAC	European Grouping of authors societies	Interested stakeholder.
LIVEUROPE	Initiative to support up-and-coming European artists through venues.	Subject to partnership agreement.

CEEMID is a pan European music data integration system based on open data, open- source software in open collaboration with the industry, statisticians and academia, best statistics, data science and AI practices. It uses many data sources about the audience, the creators of music, music works and recordings, its circulation globally and its economy. Relevance: CEEMID can transfer thousands of indicators that are reproducible and verifiable, open-source software that creates them to a European Music Observatory. In particular, CEEMID provides a useful and interesting approach to harnessing the possibilities of open data in Europe in relation to the music sector, which should be further explored by the European Music Observatory in its start-up phase. (European Commission et al. 2020, p147)

CEEMID was a predecessor of the Open Music Observatory. Although some of its data is dated, because it used reproducible research techniques in data collection and processing, some can still be updated and transferred to the Open Music Observatory. We have started this process and will continue to consult with partners about their needs or necessary approvals and permissions in the case of some datasets.

6 Standardisation of Data & Terminology

Data can only be understood with the broader concepts of information and knowledge, because data in itself is unprocessed, raw knowledge, that cannot be understood. The *EMO Feasibility Study* intuitively defines data gaps without an apparent reference to a data or conceptual model. Because standardisation is one of the key services of the envisioned European music observatory, we gave a lot of consideration to the standards to be applied, and the terminology negotiation process among the observatory's stakeholders.

In information science, a conceptualisation is an abstract, simplified view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them. Usually, when we record information about a musical work, we do not make a copy of the entire work but record some identifying properties of the work, for example, the name of its author and the name (i.e., the title), its unique ISWC identifier, and the data or registration. Composers as human beings are represented by their names, IP Names or ISNI identifiers, and date of birth and death.

A data gap can only be formally defined and filled with some reference to conceptual models of the world. A typical data problem plaguing the music sector is the amount of computer and human work needed to connect musical works and their recorded fixation, and eventually, the composers, producers, and performers linked to these objects for royalty payment. How can we define a data gap in such circumstances, and how can we fill it?

6.1 Business Processes

Since the Open Music Observatory is primarily a data dissemination hub, the definition of our services (Chapter 3) apply elements of the *Generic Statistical Business Process Model* (GSBPM), an international standard that describes and defines the set of business processes needed to produce official statistics. The GSBMP is accompanied by the *General Statistical Information Model*, which builds on the *Data Documentation Initiative* (DDI) and the *Statistical Data and Metadata eXchange* (SDMX) (Pellegrino and Grofils 2013).

The DDI and SDMX are the foundations of working with social sciences archives, statistical microdata, and processed statistical data. Their key elements are described in the Resource Description Framework of the World Wide Web and can be used in Linked Data.

The DDI-RDF Discovery Vocabulary is a draft specification of the DDI Alliance. (Hartmann et al. 2024)

6.2 Conceptual and Information Models

Data can only be understood with the broader concepts of information and knowledge, because data in itself is unprocessed, raw knowledge, that cannot be understood. The *EMO feasibility Study* intuitively defines data gaps without an apparent reference to a data or conceptual model.

In information science, a conceptualisation is an abstract, simplified view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them. Usually, when we record information about a musical work, we do not make a copy of the entire work but record some identifying properties of the work, for example, the name of its author and the name (i.e., the title), its unique ISWC identifier, and the data or registration. Composers as human beings are represented by their names, IP Names or ISNI identifiers, and date of birth and death.

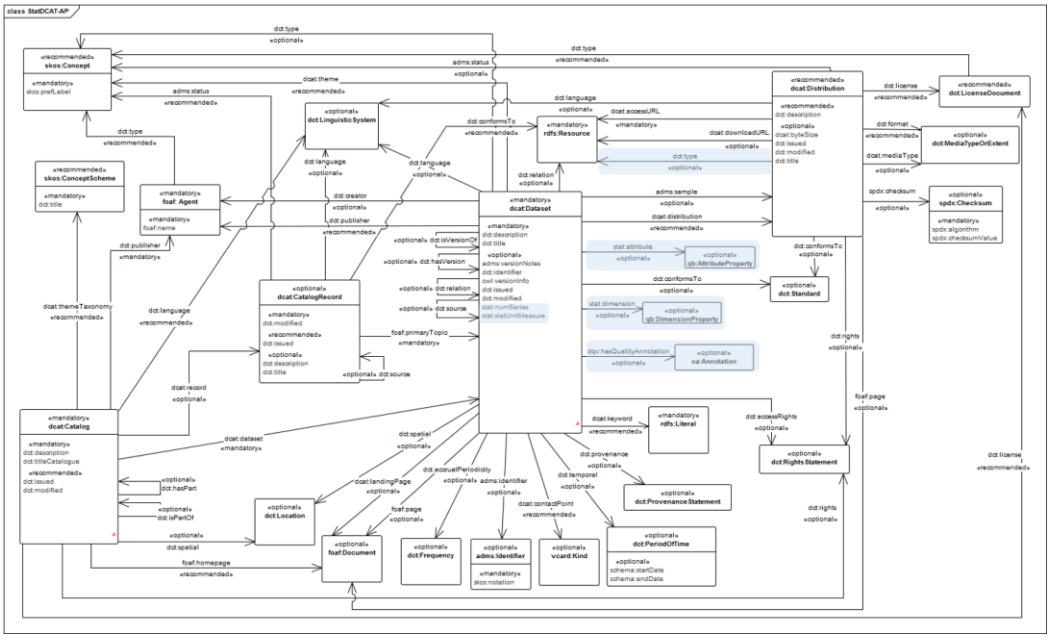
A data gap can only be formally defined and filled with some reference to conceptual models of the world. A typical data problem plaguing the music sector is the amount of computer and human work needed to connect musical works and their recorded fixation, and eventually, the composers, producers, and performers linked to these objects for royalty payment. How can we define a data gap in such circumstances, and how can we fill it?

Numerous knowledge institutions store information about musical works, as well as natural persons (humans) who composed or performed these works and contributed to their recorded fixation. If we want to inquire about composers, we must know that a composer is always a human (animals or software agents with AI algorithms cannot be entitled to composer copyrights.) We also must know that a musical work is an abstract creation, manifesting as a notation (physical or digital sheets, MIDI files) or recording (analogue or digital-physical object, or a file.) If we want to validate the composer's information connected to a recording of a particular musical work, we must access databases containing information about humans concerning some identifying properties of works or recordings.

We imagine a future European Music Observatory that is not a specialised knowledge institution and is not a library, archive, museum, or statistical agency. Instead, it should be able to consolidate knowledge from all such institutions and find ways to bring together data from private enterprises and data collection programs to fill the information gaps of the European music sector stakeholders.

Our services use the *Wikidata Data Model* as a data coordination and reconciliation model (Wikimedia Foundation n.d.). In this regard, we follow many successful EU and member-state, (Alexiev et al. 2020; Diefenbach, Wilde, and Alipio 2021; Rossenova, Duchesne, and Blümel 2022; Faraj and Micsik 2023) or music projects (Siler 2022). We particularly want to mention the excellent work of the University of Helsinki in creating WB CIDOC, a simple business process and data mapping between the Wikidata Data Model and the more complex CIDOC CRM used by extensive collection management systems (Kesäniemi, Koho, and Hyvönen 2022).

The **StatDCAT-AP** and the more general **DCAT-AP** definition of the *EU Open Data Portal* provide a bridge among library metadata systems, such as DCMI Metadata Terms (Dublin Core) for libraries, the World Wide Web DCAT standard for publishing datasets, and some core terms of the *Statistical Data and Metadata eXchange*.



Our most important reference is the DCAT-AP 3.0 specification, and its extension to statistical data by the EU Open Data Portal.

The **Europeana Data Model** (EDM) similarly provides a more straightforward connection tool among various library, museological or musical collections; it mainly builds on Dublin Core and offers equivalent classes for the more complex CIDOC CRM ("Definition of the Europeana Data Model V5.2.8" 2017). We see no problem in connecting the EDM towards RiC.

The **CIDOC Conceptual Reference Model** (CRM) provides an extensible ontology for concepts and information in cultural heritage and museum documentation (Bekiari et al. 2024).

Last, we mention some novel standards and standard candidates related to documents, microdata, and metadata documentation, such as music survey questionnaires. The **Records In Context** (RiC) 1.0 CRM and ontology were adopted in November 2023 to replace four international archival standards with backward compatibility. The DDI-Discovery vocabulary is an evolving standard that aims to describe important DDI terms with the World Wide Web standard Resource Description Framework. To keep our systems future-proof, we adopt elements of RiC and DDI-Discovery to document our question bank and codebooks (Archives Expert Group on Archival Description 2023; Hartmann et al. 2024).

> A future European Music Observatory could help with coordinating European research activities in the music sector. An EMO could also develop tools to establish cooperation between various data collection bodies. The Observatory should, therefore, also be involved in setting standards and developing common EU wide definitions that are crucial for consistency. (European Commission et al. 2020, p80)

Since the adaptation of the European Interoperability Framework and similar FAIR measures in open science, such terminological standardisation has taken place in the definition of formal ontologies, i.e., knowledge bases that software applications can use, too.

The music observatory should have competent knowledge engineers and ontologists and should be involved in the discussions of sector-agnostic ontology, for example, on the possible improvements of CIDOC or EDM, for a better representation of music.

There is also a need for the development of more usable and more widely accepted music-sector ontologies. In T5.1, we have reviewed the Polifonia Ontology Network and the Music Ontology, but we believe both have shortcomings for a full adaptation.

6.3 Identification & Entity Linking

Entity linking, also referred to as named-entity linking (NEL), named-entity disambiguation (NED), named-entity recognition and disambiguation (NERD) or named-entity normalisation (NEN) is the task of assigning a unique identity to entities (such as famous individuals, locations, or companies) mentioned in a digital resource, such as a file.

> The *MusicBrainz* free music database contains records of 20 artists named Paris (artists)↗, and 15 locations using the same name Paris (locations)↗, which all may enter a data-driven service as artists who must be credited for attribution or royalties, and as a place of an event, release, or publication. Connecting the word Paris to the correct person, group or location is the task of entity linking.

Since the inception of the world wide web, data flows across organisations and countries, and the use of local identifiers is not a good solution. International organisations of music, heritage management, science, and national organisations are increasingly shifting to the use of persistent identifiers (or permanent Identifier or handle).

> A persistent identifier (or permanent Identifier or handle), is one that never changes, so that your bookmarks and links don't break when a website or a database or an API service gets updated.

In 2024, there will be no European or international standard procedure for using PIDs, but several EU member states (Austria, Czechia, Germany, Netherlands) and other countries will have already adopted national PID strategies. Because Reprex is the current technical registrar of the Open Music Observatory, we losely follow the Dutch national strategy (Cruz and Tatum 2021) and the ID allocation practice of the *Nationaal Archief*, but this means no bias towards data partners in the Netherlands. The Dutch PID strategy does not use mandatory practices; it only recommends practices, and offers a thought-through consistent policy of using global identifiers that are not country-specific.

The structure and management of global identifiers strongly correlates with the grade of achievable automation and the potential for innovation within and across different sectors of the media industries.

Because of the prevailing problems of named entity linking, we are planning value added services to resolve named-entity recognition and disambiguation (NERD.) For this purpose, we are planning the use of AI (see Section 7.3).

6.3.1 Registers & Authority Files

Registers record every data subject belonging to a category or class: every music publisher operating in a jurisdiction, music composer with copyright claims, or statistical dataset published. Registers are essential in identifying persons and objects ("things" in information science.)

Authority files play a similar role in collections management: they provide identification information about persons or objects and tools for disambiguation. Authority files, for example, give the preferred name title for persons and musical works when available in different name or title formats, and they provide a language-independent, machine-readable identifier pointing to the correct name title. For two or more authors or performers with the same name, these identifiers help reference the proper person (or object.)

Registers are valuable and indispensable for many digital workflows. They serve as the foundation of various processes, such as statistical sampling (determining who should receive a questionnaire) or

copyright management (deciding who should receive the royalty payment). Their absence or inefficiency can significantly hamper these operations.

Unfortunately, the music industry has long missed access to reliable, open registers. The reasons for this are beyond the scope of this report, but we highlight that the underlying reasons for closed and not interoperable registers are deeply rooted in the conflicts of interests among different sub-sectors of music and are unlikely to be solved in a short time. Therefore, music enterprises, researchers, professionals, and curators will need identification services and identity brokerage services for a long time.

Creating and maintaining high-quality registers require significant professional and financial commitments, and they can form a vital service of a future European Music Observatory. Currently, we are experimenting with three service levels in the Open Music Observatory.

- We create our own transparent and interoperable identifiers within the OMO for persons and their groups (ensembles, bands, orchestras, associations…), legal persons (music businesses, collective rights management agencies, …), events (recording, composing, performing events, festivals, conferences, …), musical works and their manifestation (books, works, recordings, sheets.)
- We create integrity brokerage services and middle-term identification via Wikibase and Wikidata. Our identifiers are connected to middle-term Wikidata and Wikibase QIDs, which also serve as graph nodes to registry, library, collections, and industry-specific identifiers.
- We are piloting data improvement services that can find erroneous identifiers or add correct identifiers to various datasets.

6.3.2 Open and persistent identifiers

In line with the practice of the Netherlands, we prefer the use of the following identifiers:

ISNI: preferred persistent identifier for names of people and groups. The use of ISNI is also preferred by Apple Music, Spotify, and as a pilot it was introduced by Teosto, the Finnish national collective management society; it is being considered in many use cases for adoption in all CISAC societies. ISNI is the ISO certified global standard number for identifying the millions of contributors to creative works and those active in their distribution. (Camp, Lieber, and IFLA 2022)

For legal persons, we are discussing the terms to use the OpenCorporates ID, because many organisations at this point do not have an ISNI.

ORCiD: preferred persistent identifiers for music researchers and scholars. This is in line with the Horizon Europe and the European Open Science Cloud recommendations; ORCiD itself only adds functionality to ISNI; i.e. each ORCiD ID is at the same time registered as an ISNI.

VIAF: VIAF is the shared authority file of national libraries. It offers more services than ISNI and includes an ISNI for the author.

DOI: we use the Digital Object Identifier for publicly released documents.

ISBN: We issue ISBN identifiers for long-form publications of our partners. (ISO 2017b)

6.3.3 Not open, music-industry specific identifiers

Book and music sheet publishing uses the ISBN and ISWN, professional and magazines and scholarly music journals use the ISSN, and the music rights management uses ISRC and ISWC. These standards usually resolve an identifier to some network location where metadata or the object itself can be found. There are many advantages and disadvantages of this model.

For example, the ISWC identification of musical works is the backbone of copyright management, and it is a closed and consistent system developed over many decades by the member organisations of CISAC. The downside of this closed system is that the metadata about the works identified by ISWC is strictly available only to CISAC member societies. While CISAC offers an API for the individual lookup of ISWC for one example of a musical work, currently it does not allow bulk access to the registered data.

We have already started a discussion with some music industry registers about connecting the Open Music Observatory to their systems. We are planning to present our proposals on the CISAC Good Governance seminar to be held in December 2024.

Musical works ISWC: nternational Standard Musical Work Code is a unique identifier for musical works. It is adopted as international standard ISO 15707 (ISO 2022). OpenCollectons ID: Our ID for music works (only if we publish data about them.)

Sound recordings ISRC: The International Standard Recording Code (ISRC) is the international identification system for sound recordings and music video recordings. (ISO 2019; Authority 2021) OpenCollectons ID: Our ID for sound recordings (only if we publish data about them.)

Music sheets ISWN: The International Standard Music Number currently identifies published music sheets (ISO 2022).

ISBN-13: Before the introduction of ISWN, published sheets were identified by the ISBN book identifier. ISBN-10: The older format of the ISBN book identifier, which predates both the ISWN and the 13-digit ISBN used to identify music sheets. ISCC: The International Standard Content Code (ISCC) is an identifier for numerous types of digital assets. This is our preferred identifier for not published sheets. (ISO 2017b)

For unpublished works, our preference is the use of the brand-new ISO-standard ISCC because it was designed precisely for the use case we were looking for. It is free to generate, generated from digital content (or its digital copy), and can connect various local or lesser-used identifiers.

Datasets DOI: DOIs are assigned to each distribution of a dataset. As datasets are often continuously filled, these datasets will have periodic versions with versioned DOIs (from Zenodo.) OpenCollectons ID: Our ID for unversioned (continous) datasets, pointing to the latest available version of the data.

Codebooks URI: Whenever possible, we use standard codebooks of SDMX or Eurostat, and provide a URI to the codebook, and provide dereferencing to the codebook definition. OpenCollectons ID: Our ID for our codebooks, regardless if they are same as the SDMX/Eurostat standards, or we create a non-standard coding for a novel dataset.

Questionbank URI: Whenever possible, we use standard questionnaires, and provide a URI to the codebook, and provide dereferencing to the DDI questionnaire item definitions. OpenCollectons ID: Our ID for questionbank items.

6.3.4 Lyrics

In many genres, lyrics are very important parts of a musical work, and there is a growing demand and need to provide or analyse the lyrics of the work. For example, in our X, we want to create location-aware music services and encourage the public performance of music made in Bratislava or music somehow specific to Bratislava within the public places or radio stations of Bratislava. One possible semantic connection to this environment is that a song is about Bratislava (Berlin, Paris, or Germany.)

Access to the lyrics part of the music is not straightforward, mainly because the lyrics may be arranged from a literary work. We see lyrics identification and semantic analysis as the next immediate step to our location-aware application, for which we are looking for good industry solutions. In many cases, we will likely need to rely on the ISCC code as a temporary identifier for lyrics databases that were not available in a licensed format earlier.

6.3.5 ISCC

The *Open Music Observatory* will start to implement the newest ISO-standard open identifier, the ISCC-CODE. ISCC is inverting the principle of a centralised register. It generates the ISCC code from the digital content object itself, therefore no third-party lookup is needed for finding the identifier of the object.

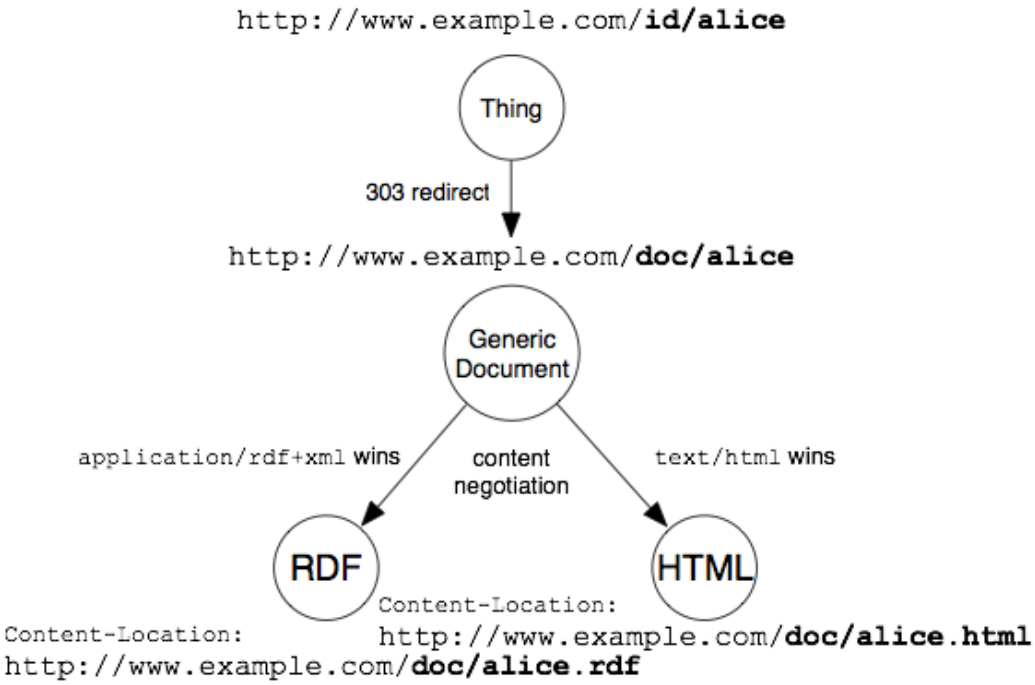
ISCC registration becomes necessary when an ISCC code needs to be globally unique, publicly discoverable, resolvable, owned or authenticated. While these features inevitably require some kind of registry, not all of them require a centralised institutional registry. The ISCC specifies the necessary protocols to implement the aforementioned features in a decentralised, federated environment and across multiple public blockchains. Given a registered ISCC code, an application can unambiguously determine on what blockchain (if any), by which account, and at what time an ISCC has been registered.

Registered ISCC codes refer to an authoritative public blockchain network. This indicator is part of the ISCC Code itself, such that codes registered on different networks cannot collide. This guarantees uniqueness of ISCC codes across multiple blockchains. Ownership of ISCC codes (not the identified content) is granted to the signatory of the first transaction for a given ISCC code on the corresponding blockchain.

As such the ISCC fulfils a distinct role and is not a replacement for established identifiers. Rather it is designed as an umbrella standard to augment established identifiers with enhanced algorithmic features. It can be used in the metadata of existing standards or support discoverability (reverse lookup). We will use for precisely this application: whenever we receive content that is not identified by a DOI, ISNI, ISWC, ISRC, or other standard identifier, we will assign an OMO identifier and enhance it with the ISCC features. This will help later linking to the preferred global, persistent identifiers.

6.3.6 OMO Identifiers

We create our own identifiers for persons and things. We follow the practice of the Dutch national archives in the creation of PIDs, and we make them URIs following the W3C recommendation.



music.dataobservatory.eu/{type}/{concept}/{reference}

For {type} we utilise the following definitions: - {id}: an identifier {type} for dereferenced identifiers.

- {doc}: a documentation {type} for the documentation of persons and objects.
- {def}: a definition {type} for ontologies.

For {concept} we utilise three categories:

- music.dataobservatory.eu/{id}/{person}/{reference} for persons, in order to synchronise with national and international name spaces.
- music.dataobservatory.eu/{id}/{place}/{reference} for places, in order to synchronise with national and international name spaces.
- music.dataobservatory.eu/{id}/{oc}/{reference} an other objects, such as musical work, a sound recording, a group a persons.

7 Data Improvement & Innovation

The music sector was one of the early adopters of digitisation and is a highly data-driven sector of the economy. Because it relies on data, business and public policy problems often accompany data problems. In the previous section, we have shown how we aim to increase the data available for the sector. Now, we focus on improving the data's quality and usability.

7.1 Value-Added Data Services

7.1.1 Data Sharing

"Data sharing" means securely sharing data among parties who do not want to expose their data to third parties or protect the personal data in the datasets. Agreeing and organising data sharing legally, semantically, syntactically, and technically can be challenging. This is the role of the Open Music Dataspace behind our observatory.

The data sharing infrastructure behind our dataspace is the Reprexbase system, which is an extension of the Wikibase system using various open-source (and, in a few cases, non-open-source) software components to connect the Wikibase system with music sector databases and data sources.

Given the sensitive nature of the data we handle, including business confidential and GDPR-protected data, we maintain strict segregation. Data batches from stakeholders are kept in separate instances and are integrated only after thorough review by the data protection officer and curatorial team, ensuring the highest level of data security.

In our Slovak prototype, we keep SOZA's data in an insulated instance because the copyright management organisation must not release GDPR-protected and business-confidential information. Some of the data needed for NERD operations or the establishment of the Slovak Comprehensive Music Database is then sent to a joint instance concerning those data subjects (i.e. authors or their heirs) who agree with our data handling. This is where they meet public catalogue and database data from public libraries, open knowledge graphs, and the Slovak Music Center.

The data that should be made public is then further exported to Wikibase Cloud, where it becomes public and available for all stakeholders. From Wikibase, it is also synchronised with Wikidata, the world's largest open knowledge graph.

7.1.2 Fix-the-data

"Fix-the-data" means improving the data quality by finding or imputing missing values or finding and replacing erroneous data entries. In terms of metadata, adding further machine-actionable information to already existing datasets can improve their usability.



Our fix-the-data services do not increase the size of the data available to our partners, but it increases the quality of their datasets or databases.

7.1.3 Data Linking

"Data linking", data fusion, or data matching means correctly joining data from different datasets (data sources.) Many fix-the-data problems initially arise from imperfect data linking, for example, mistakes in currency rates, units of measures, coding of geographical entities, misplaced decimal delimiters on

the level of data, or misunderstandings of the meaning of “artist income” or “popularity score”, or other non-self evident variables. An even more subtle problem is joining data from two questionnaire surveys created with different sampling algorithms and different standard (measurement) errors.

Data linking or data fusion is a way to join many small databases into a large, federated dataset. This way, relatively small music organisatiosn can benefit from access to big data.

7.1.4 Registration services

In Open Music Europe, other tasks deal with the policy problem plaguing the music industry: even though it needs access to an exceptionally high number of registers (due to the fragmentation of the copyright and several neighbouring rights), access to such registers is limited or impossible. Often, the registers carry legacy problems that make them less functional in trustworthy data and AI systems.

“A register aims to be a complete list of the objects in a specific group of objects or population.” (Anders and Britt 2007). Statistical data collection and rights management are just two service areas whose workflows depend on well-functioning and accessible registers. The statistical business register is an essential tool for creating survey frames or sample frames, in other words, to organise statistical data collection. A copyright or neighbouring right register is necessary to organise royalty collection.

A statistical register is necessary to decide who should get a data request:

- For a *sample survey*, the register is used to draw a lottery of population members who will be invited to provide data.
- In a *census*-type survey, all registered members of the population, for example, all music labels, will receive an invitation to an interview or form.
- In the case of *a register-based survey*, all members of the register, for example, all collective management societies in the territory, will be requested to send data directly from their databases.

In other work packages of the Open Music Europe project, we are experimenting with statistical data coordination among the music sector and statistical authorities. Without recalling the details here, as digitisation exponentially increases the amount of structured data in the private sector, it is a growing trend in statistical innovation to rely on data held by the private sector to make more granular or timely official statistics. For consumer spending statistics, costly and imprecise surveys of randomly selected citizens putting their purchases in a diary, some statistical authorities directly process data from cash registers or credit card spending. We envision a similar statistical collaboration among statistical offices and collective rights management organisations because it is easier to report music royalty accounts than to ask musicians to talk about their complex income streams in interviews or on questionnaires.

We see the role of the Open Music Observatory in providing a methodology and digital data infrastructure for such statistical collaboration. In other work package tasks, SOZA and Reprex will create so-called satellite business registers to harmonise the data collection of the observatory with the Slovak statistical authority. More about this work: (Antal 2023)

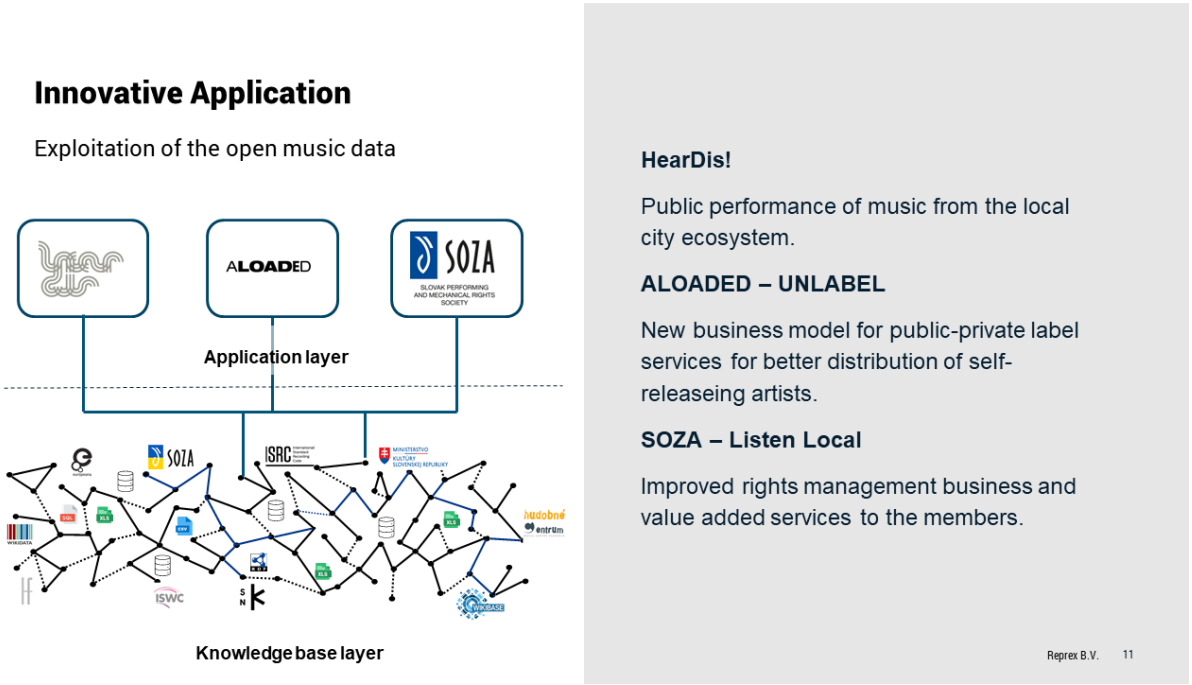
Such services, similar to data linking and some new services that will build on the data and the data API of the Open Music Observatory, rely on the provision of technical services for registration. The Open Music Observatory has its register, too.

Registration is a costly data service with vast economies of scale, so providing more affordable registration services for the European music sector could be an important service.

- ☒ In our piloting phase, we rely on cooperation with the Slovak National Library to test the usability of the VIAF authority file system for identifying names.
- ☒ Our dataset distributions use the DOIs from Zenodo, which also provides our long-term archive.
- ☒ For certain assets, mainly photographs and scanned documents, we rely on the new ISCC registration, a long-term solution for some music industry applications.
- ☒ Reprex registered an imprint, the Digital Music Observatory, to place long-form publications as books into library systems.
- ☒ We are investigating the costs and benefits of finding an ISNI registrar partner or creating a roadmap for making the Open Music Observatory a registrar itself.

7.2 Use Cases

In 2023 the Open Music Europe project applied for the Module A of the *Horizon Results Booster* (HRB) provided by Trust-IT Services↗. The HRB aims to provide a tangible contribution to the dissemination of results and recommendations of research projects related to the European Commission Priority areas.



7.2.1 Data Health Services for Collective Management

Entity linking and data linking are among the biggest technical problems in rights management. Because music authors, producers, and performers have three royalty streams and do not share an

interoperable registry, the connection of musical works (compositions, ideally identified by an ISWC code), their sound recording manifestations (identified on all digital services with and ISRC code), and the various identifiers of performers require costly manual and technical identification.

There are numerous projects underway in the music industry to resolve this problem going forward. In the United Kingdom, PRS's Nexus programme↗ is developing a solution with the provisioning of preliminary ISWC registration to keep the recording and composition connected from the birth of a new recording.

The Open Music Europe project, on the other hand, is pioneering a different route for already existing sound recordings, with the linking of public sector catalogues of heritage and library collections with rights management information; particularly with relying on the VIAF shared authority files. SOZA and Reprex are expected to present their MVP on the CISAC Good Governance seminar in December 2025.

Modern registers typically assign a unique identifier, known as a URI, to their data subjects (our registered objects). A 'Cool URI', which resembles a URL, offers a practical advantage. When used as a URL, it generates a human-readable HTML file about the registered person or object. This can be particularly useful when processed by a graph application, as it provides crucial information about this person or object in a machine-readable (XML, JSON, TTL, or NQUAD) file.

For example, the VIAF identifier number 89006617 can be placed into the http://viaf.org/viaf/89006617 URL, which provides as access to the cataloging information of works created by, or written about the great etnomusicologist and modern composer, *Béla Bartók*.

Bartók, Béla, 1881-1945. **Bartók, Béla**

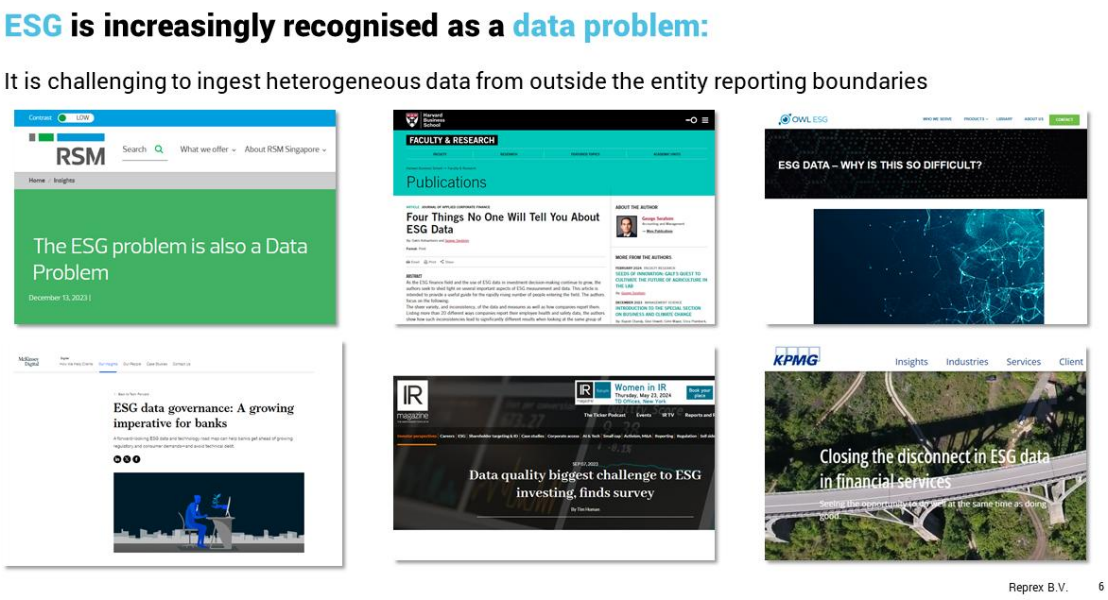
Modern platforms, such as Spotify, use similar identifiers. For example, the Spotify Artist ID 2fIUlieTjLTaNQUIKHX5B8 resolves to Celeste Buckingham's available recordings on the platform via the URL https://open.spotify.com/artist/2fIUlieTjLTaNQUIKHX5B8.

The problem is that music creators are often present on more than 200 digital platforms, each of which has its identifier policy and requires the repeated import of the artists', works', and recordings' data. To consistently report such metadata is costly and complex, even for major labels and publishers with a dedicated IT system. No wonder we saw before our project in our own *Feasibility study* that more than 50% of artist data needed fixing on digital platforms.

Relying on many local identifiers on otherwise interconnected computer systems will always create a costly and error-prone data exchange. Unfortunately, the music industry has never agreed to use genuinely open, high-quality registers. These changes were made during the period of our project. For example, large platforms like Apple, Spotify, and some collective rights management organisations started using the ISO-standard name identifier (ISNI) to avoid the high prevalence of multiple same-name persons and musical groups. This transition is yet to begin, and it is incomplete, so the music sector will likely need to invest large IT resources into entity resolution in the next decade.

7.2.2 Sustainability Reporting for Music Organisations

The Music Innovation Hub and Reprex will develop a CSRD-compliant sustainability reporting tool in 2024-2025. The reporting tool aims to provide an accurate and affordable ESG reporting facility that follows the European ESRS standards for music enterprises that create their financial reports according to the simplified reporting rules allowed by member states for microenterprises.



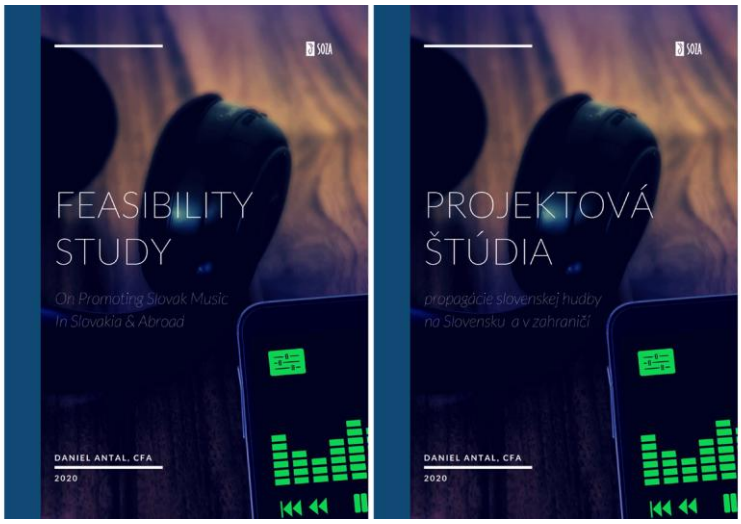
More than 95% of European music enterprises (in some member states, this reaches 100%) apply simplified financial reporting. For such companies, there are no CSRD-compliant ESG reporting tools.

We identify the reason for this market failure as follows:

- ☐ The CSRD Directive imposes the responsibility of connected financial sustainability reporting on large and public companies and applies it to their entire value chain. The music industry lacks such large enterprises that would have taken a piloting role or played a pivotal role in establishing the standards.
- ☐ Music enterprises and their trade associations do not act proactively because they believe they must follow the data provision instructions of the directly affected B2B buyers, financiers, or corporate sponsors.
- ☐ The standardisation body EFRAG has de-prioritised the cultural and creative industries in setting industry-specific standards favouring sectors with a much higher adverse environmental impact.
- ☐ While small music businesses do not feel a compliance push, as they are not directly responsible for applying the ESRS, they also miss out on the opportunities provided by green financing and insurance.
- ☒ MiH and Reprex will pilot a service suitable for microenterprises, reducing compliance costs from 1500 euros to 500 euros per entity.
- ☒ This new application will rely on the Open Music Observatory's Music Economy and Sustainability pillars and will derive its benchmarks, science-based targets and coefficients, and input-output tables.

The MVP of this service was developed with a MusicAIRE microgrant, and it is the project's background. A scale-up will be demonstrated with the use the Open Music Observatory's open data API.

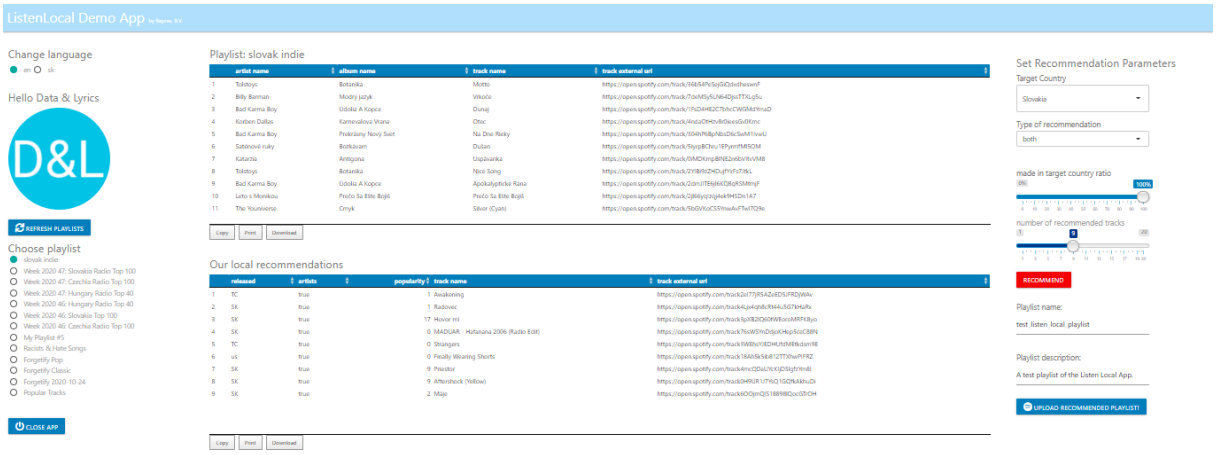
7.2.3 Listen Local



The Feasibility Study On Promoting Slovak Music in Slovakia And Abroad is an important background of our project.

In 2020, with a microgrant from the Slovak Arts Council, we created a Feasibility Study and a demo application called Listen Local (Antal 2020b). The study examined why the Spotify algorithm struggled to recommend Slovak music within Slovakia for Slovak people. We also created a demo application that modified the user's Spotify recommendations to voluntarily comply with the local content guidelines applicable to local radio stations. The user could also listen to a lower or higher percentage of regional works.

Our critical finding was the very pool data coverage and quality of the Slovak repertoire, which is mainly sent to distribution without the professional assistance of a commercial music label. Self-releasing artists and micro labels do not have the necessary metadata know-how, IT and data specialists to prepare their new releases for algorithmic curation by recommender engines of digital streaming platforms, radio stations, or large festivals.



Our conceptual demo application was able to make recommendations on voluntarily meeting the local content guidelines, but it was only supported by a relatively small Slovak Demo Music Database, and could only work with Spotify, which has the most transparent and open API of all streaming providers licensed to the territory of the Slovak Republic.

We aim to develop applications to create a local content-aware public performance music stream.

- ☒ HearDis! aims to integrate such location-aware metadata into its background music playlisting service.
- ☒ We are planning Listen Local applications for radio stations to voluntarily review their current playlists for compliance with local content regulations and, if they fail to reach the statutory local content quotas, to recommend suitable recordings to their playlists.
- ☒ The OMO will disseminate the necessary data for these new services.
- ☐ The data is not yet available, as the creation of the Slovak Comprehensive Music Database is a task of its own that will be ready by November 2025 in WP2.

7.2.4 Unlabel

Unlabel is a planned service aimed at self-releasing artists and micro labels that need a functional data/IT department. Therefore, they are at a disadvantage compared to significant independent and major releases because they usually need to meet the high documentation standards necessary for a successful digital distribution strategy and engagement with algorithmic curation of streaming-, radio-, or festival playlists.

Self-releasing artists and micro labels bring ill-documented new content to digital distributors like ALOADED. Digital distributors must maintain an arm's length standard for all labels, small or large, independent or major. ALOADED or other distributors cannot cross-finance the data problems of self-releasing and micro-label artists from the client revenues of more prominent labels.

We identify the problem as a market failure and a technical failure:

- ☐ In some developing markets, insufficient royalty revenues do not allow the presence or professionalisation of record labels with an IT and data management function because the payment of IT or data specialists or to keep external suppliers at least on a retainer cannot be financed from the label-artist revenue split.
- ☐ Manual metadata provision without metadata specialists and tools leads to inferior data quality. Our feasibility study has shown that more than 50% of the releases have data shortcomings, and 17% have poor data representations that make algorithmic recommendations for these releases impossible. This creates a vicious circle because poor data quality translates into low visibility, low usage of such repertoire, and, therefore, low income. The cost of data improvement has no sustainable financial basis.
- ☒ In 2025, ALOADED, Reprex, Slovak Music Center and SOZA will conceptualise and plan a new public-private business model that aims at those rightsholders who do not have a *technically* proper label representation as a substitute for non-available market services.

Our planned "Unlabel" service will provide documentation and metadata improvement services for self-releasing artists. This service, similar to current white-label services, will strictly address market failures and not compete with label services. We aim to provide a necessary level of data consolidation and improvement so that these artists can have equal opportunities in digital distribution services.

The service will be connected to the *Slovak Music Dataspace* and its Slovak Comprehensive Music Database. We will provide a PPP business model for the onboarding and proper documentation of self-

releasing artists on a large scale and the efficient, API-based provision of their digital distributor. Aloaded will provide the distribution services, Reprex will provide the data services, and SOZA and the Slovak Music Center will work out the details of minimal customer service for such labels.

7.3 Use of AI Systems

For the entity linking, related to our planned value added Section 7.2.1, we are planning to use in the future AI algorithms, particularly inference engines. The main goal of the system is to help matching correctly named entities, particularly rightsholders, musical works and recordings. The system is not yet in place. An adequate description will be provided for overview and will be brought to the attention of the Ethics Advisor during the upcoming meeting of the Ethics Board.

We cannot provide a full risk assessment because the service is not planned in detail yet. However, our preliminary risk assessment suggests **low levels of risk**, partly, because we plan to deploy AI in music/culture, which as a domain not seen as a high-risk area by the European regulation, and partly, because our system will not autonomous, will retain human-in-control, and will not influence the decisions or anyhow engage with end-users.

We were conscious of the potential risk involved, and both the control structure and the data governance were planned over the course of 10 months.

- ☐ Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society? **No**. The system will only help qualified persons in rights management to faster and more efficiently preview potentially unlinked entities.
- ☐ Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way? **No**. The system in no way is considered as an end-user system.
- ☒ Please determine whether the AI system (choose as many as appropriate) overseen by a human: Is overseen by a Human-in-Command.
- ☒ Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight? **Yes**. The system is not making autonomous decisions.
- ☒ Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? **Yes**.
- ☒ Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent? **Yes**.
- ☒ Data Protection Impact Assessment (DPIA) **Yes**.
- ☒ Designate a Data Protection Officer (DPO)24 and include them at an early state in the development, procurement or use phase of the AI system? **Yes**.
- ☒ Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications)? **Yes**.
- ☐ Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation)? **Not applicable for NERD**. The aim of the application is to

detect errors in name attribution and to protect the moral and economic rights of the (named) rightsholders.

- ☒ Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system? **Yes**.
- ☒ Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle? **Yes**.
- ☒ Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data? **Yes**.

We do not consider that the system has wider risks or negative impacts. The algorithm is designed to cure sources of data biases that result in a late or missed payment for some rightsholders.

8 Conclusions & Next Steps: Towards a European Music Observatory

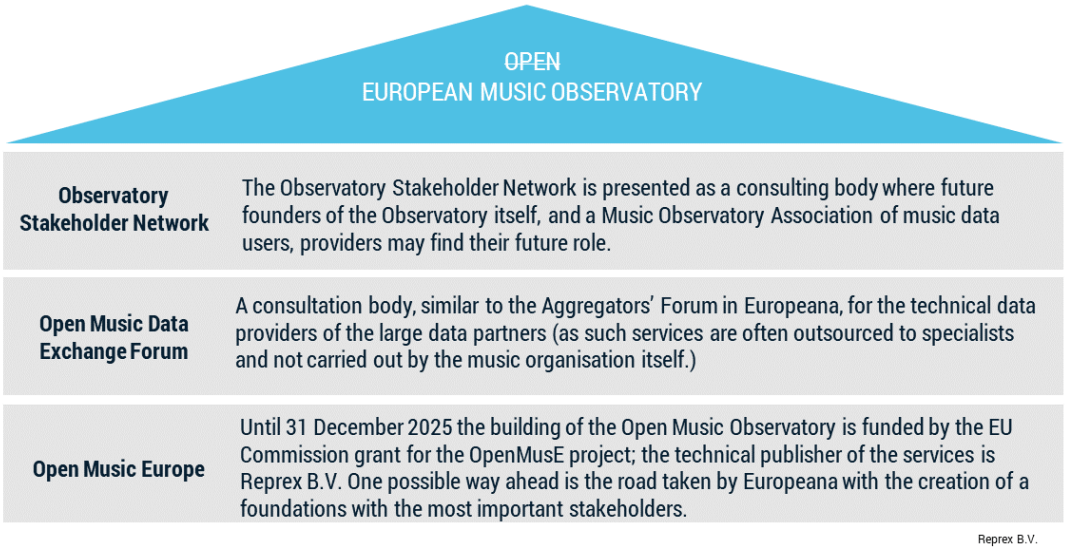
In this report, we introduced the Open Music Observatory, a complex data curation and information management system with a novel observatory governance model. The system is technically ready and has been tested with real-life use cases and scenarios: it handles complex data extract-transform-load processes, GDRP and confidentiality problems, and many aspects needed to make a data (sharing) space work. Because it is based on open-source components and on open data standards, it is set on an optimal path to grow as a data ecosystem.

The next development phase of this observatory will be critical: we must persuade music sector stakeholders to trust and use our system for data sharing well beyond the stakeholders of the Open Music Europe project. Once we reach out and onboard external partners, we will certainly have to modify business processes and permissions and provide further technical help, software components, and manuals. Besides these obvious iteration steps, we would also like to agree on a governance model that allows the extensions and the ownership transfer of this Open Music Observatory to a truly European, shared, public-private observatory that can serve the music sector for a long time.

8.1 Co-Creating a Governance Model for the Observatory

Our current roadmap for the institutionalisation of the Observatory is based on the example of *Europeana*, a joint collection of European libraries and museums. Europeana 1.0 was built over 2.5 years in the *European Digital Library Network* project. It remained a decentralised network of three organisations: the Europeana Foundation, which is the technical operator of the digital collection; the *Europeana Network Association*, which is free to join for all interested parties; and the Aggregators' Forum, which is a technical coordination body to help those technical data providers that send or exchange data with Europeana.

To draw on this analogy, the Open Music Observatory is being created by the Open Music Europe project. As stated in our Grant Agreement and the connecting Consortium Agreement, we treat the duration of the Open Music Europe project as a prototyping and development phase when multiple institutionalisation forms are still possible. Based on our Agreements, we formed the Observatory Stakeholder Network as a stakeholder group to set priorities and express opinions on our work and the potential longer-term institutionalisation alternatives.



We consulted several stakeholders throughout the project, and eventually, following their advice, we decided to form this advisory council when we already had a working data dissemination infrastructure and reviewable data in it to start their work. Not prejudicing any later presented organisational proposals, for the time being we follow functionally the organisation of Europeana, because we need to deal with similar problems.

Europeana started out as a project of a few national libraries and after 16 years of existence, it aggregates digital collections from more than 3000 organisations of Europe. Such a large-scale but decentralised organisation requires a multi-tier, multi-functional governance structure, where large stakeholders can take ownership, smaller stakeholders can democratically participate and cooperate, and technical providers can work separately on technical-only problems.

- ☒ The **Observatory Stakeholder Network** is presented as a consulting body where future founders of the Observatory itself, and a Music Observatory Association of music data users, providers may find their future role. We hope that some of the initial members will become founders in the future European Music Observatory, and others will be forming a democratic Music Observatory Association for grassroot aggregation and dissemination. (Observatory Stakeholder Network↗).
- ☒ Our **Open Music Data Exchange Forum** is a consultation body, similar to the Aggregators' Forum in Europeana, for large data partners.
- ☒ As stated in our Grant Agreement, we will offer various institutionalisation proposals by the end of our project. If the representative stakeholders of the European music sector will choose not to create a European Music Observatory, we will continue to operate the Open Music Observatory with any interested party.

8.1.1 Observatory Stakeholder Network

The Observatory Stakeholder Network is a temporary organisation defined by the Open Music Europe project Consortium Agreement. It is an observatory volunteer advisory body.

The Europeana Network Association (ENA) is a strong and democratic community of experts working in the field of digital cultural heritage. We are united by a shared mission to expand and improve access to Europe's digital cultural heritage. The Association is free to join and we encourage our members to get involved and benefit from all the ENA has to offer.

We will invite every organisation that has promised data for a future European Music Observatory or has shown interest in the previous CEEMID or Digital Music Observatory collaborations to join our network. In the longer term, we envision some pan-European representative and umbrella organisations becoming founders or board members in the more permanent observatory organisation. For individuals, research groups, and micro-enterprises, we will suggest setting up the *European Music Observatory Network Association* (or similar entity), learning from the experience of Europeana.

8.1.2 Open Music Data Exchange

Aggregators work with cultural heritage institutions to gather authentic, trustworthy and robust data and make it accessible through Europeana. All Europeana aggregators are members of the Europeana Aggregators' Forum (EAF), a network of national, regional, domain and thematic aggregators who - among others - work to exchange the knowledge and best practice that supports aggregation and data sharing with Europeana.

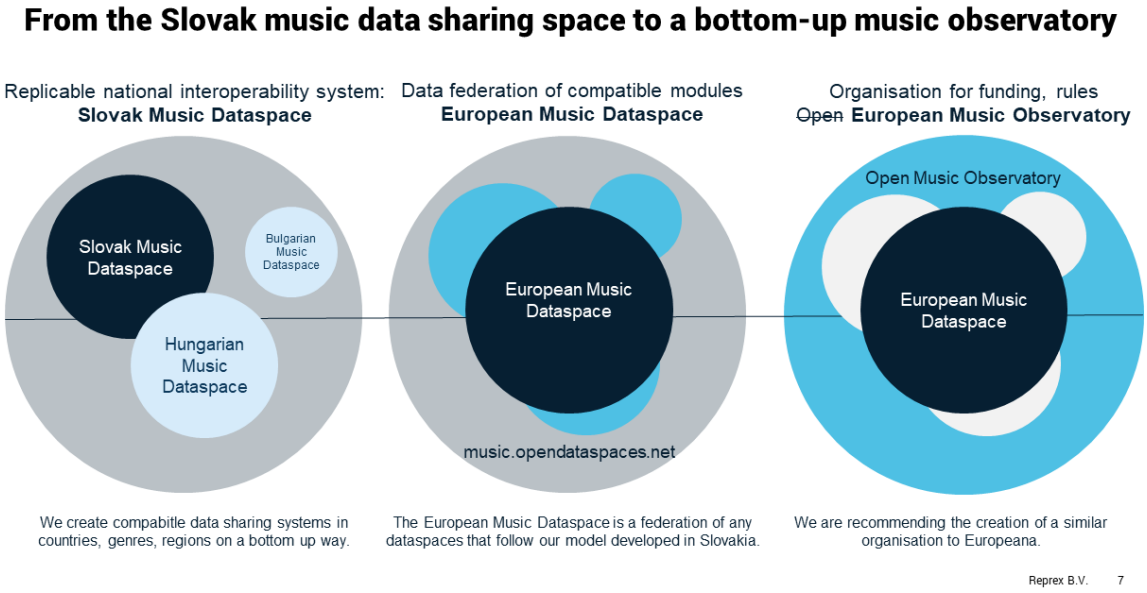
The Europeana Aggregators'' Forum held in May 2024, comprises 13 cross-national aggregators and 30 national aggregators who help cultural heritage organisations aggregate their collection datasets into the Europeana.

We imagine a similar organisation comprising companies that are not data owners but provide data and IT services for important national or pan-European stakeholders. While they cannot give permission or license data to the Open Music Observatory, their expertise in carrying out data integration will be invaluable in making our data-sharing space behind the observatory work.

8.2 Bottom-Up Expansion of the Observatory in a Federation Model

A data federation is collection a software process that allows multiple databases to function as one. This virtual database takes data from a range of sources and converts them all to a common model. This provides a single source of data for front-end applications. Data federation is a key feature of the data (sharing) space. Data federation can only occur after the legal conditions, business process alignment and semantic harmonisation allows the data to be legally and meaningfully integrated.

As explained in Section 2.3, we have started the prototyping of our governance and interoperability model in the Slovak Republic.



The data integration that we started in the Slovak Repubic can be replicated in other parts of the European music sector, among any two or more data sharing partners.

This is an essential aspect of our bottom-up approach: similar data integration can start in other parts of the European music sector. It can take shape in another national (or regional) data-sharing space, a genre-specific one (for contemporary classical music, which has a relatively distinct ecosystem from popular music), or in a business subsector, like publishing. As long as these stakeholder groups use the same standards and architecture, we can gradually connect them into a European music dataspace. We will add such replication cases outside of Slovakia in the Open Music Europe project. Because our system is shaped as a decentralised graph without a central node, it does not matter which countries, sub-sectors, or genres follow Slovakia; the country has no unique role in the eventual Observatory, apart from being a pioneer and knowledge centre in the music data exchange.

The data (sharing) space is not only a national data exchange: it allows us to coordinate data with open knowledge bases and their data graphs internationally; we will demonstrate whom we exchange data with Wikipedia/Wikidata, CIDOC-compliant collections, RiC-compliant archives, or SDMX-compliant statistical services.

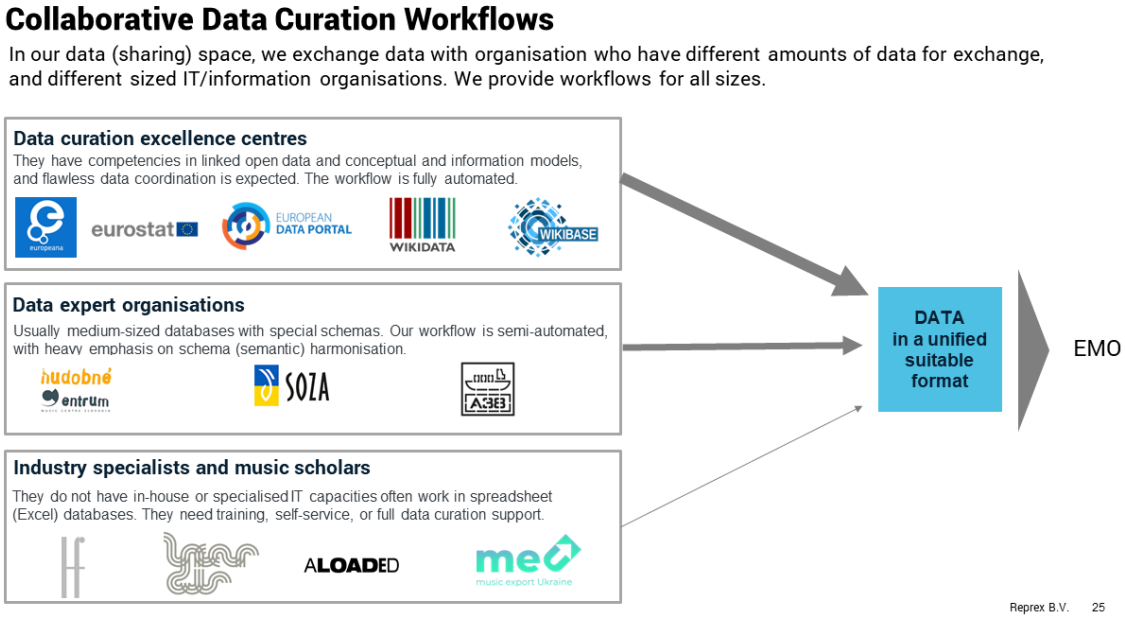
8.3 Practical Next Steps in the Project

Task 5.1 developing and validating the Open Music Observatory is running until the end of the Open Music Europe project, and it is connected to almost all other project tasks. Methodological and research tasks of the project identify data gaps and methodologies to fill those gaps, while other tasks provide us with more direct input in the form of data that can be disseminated via the Observatory.

With the current deliverable the Open Music Observatory has been released with a governance model to combine existing open data sources and privately held music sector data with data created within other project tasks. This marks a milestone, linked to MS4 of the project for this system development, as the Observatory is now ready for use by all internal and external stakeholders, enabling open data collection and real-time reporting to begin.

To begin the validation and improvement of the Observatory we have imported data from the project's background or other sources and encouraged project partners to work with our evolving system already before this milestone. As a result, our observatory systems are not empty at the start, but they also lack the critical mass of data to train data improvement algorithms or provide value-added services. Those functionalities will be utilised as other project tasks report their results and hand over their data to us; and as external stakeholders start to trust us with their data.

For the long-term sustainability of the observatory and for a successful exploitation, we need to bring as many stakeholders as possible into our data (sharing) space and our Observatory Stakeholder Network.



Different level of service and help for data exchange with stakeholders

We have done extensive system requirement solicitation and ingested sensitive or complex data as we were developing the system. Based on almost 1.5 years of experience, we believe that we need to further refine the auxiliary tools for three data curation/exchange scenarios.

- ☒ For data curation excellence centres, like the EU Open Data Portal, Europeana, Wikibase, and Eurostat, we must ensure that our machine-to-machine communication works better and exchanges richer metadata. We improve fully automated and well-standardised workflows.
- ☒ For larger music sector stakeholders who have an advanced internal IT function and often some library/information science competence but no experience in machine-to-machine communication with standard or open protocols, we need to provide servicing and help with semantic schema adjustments, as they are usually able to export data in tidy and well-organised format. We expect collective rights management organisation, national libraries or larger specialised public music libraries, archives, museums, and larger music information centres to fall into this category.
- ☒ For organisations that do not have an internal IT/data/library function, or work with a limited, generalist IT function, we need to provide simple tools. For simpler, spreadsheet-based

data sources, we would like to improve tutorials and tools for self-servicing. We need to take a cost/value assessment for small organisations with big and complex databases because they will likely need knowledge or data engineering help, which is a limited resource in our project. We expect small enterprises, NGOs, and trade associations or umbrella organisations to fall into this category.

To reach our objectives related to the Open Music Observatory ,we will take the following steps after this milestone:

- ☒ We introduce the system to those organisations that showed a willingness to share data with a future European observatory (See: Section 5.2), and ask for their feedback and at least test data samples;
- ☒ Based on the external data samples, in our supporting task, we will develop further extract-transform-load components to ensure a smooth data export/import according to the needs of the representative European music stakeholders;
- ☒ We will disseminate and improve with feedback our data sharing space model, and particularly the accompanying governance model, with experts in the field of data governance, data sharing, and various music industry data and metadata interest groups;
- ☒ We will finalise the creation of the Slovak Comprehensive Music Database, enabling the MVP demonstrations of several value-added services, as described in Section 7.1. We plan to start disseminating these use cases on large stakeholder forums in November 2025 in the hope of finding additional replication partners and users.
- ☒ An adequate description of our planned use of AI will be provided for overview, and will be brought to the attention of the Ethics Advisor during the upcoming meeting of the Ethics Board (see Section 7.3.)
- ☒ We will ingest a large batch of survey data from other work packages in 2025 Q1.

© 2024 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295

References

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley, eds. 2020. "Data Catalog Vocabulary (DCAT) - Version 2." W3C. https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/.

Alexiev, Vladimir, Plamen Tarkalanov, Nikola Georgiev, and Lilia Pavlova. 2020. "Bulgarian Icons in Wikidata and EDM." *Digital Presentation and Preservation of Cultural and Scientific Heritage* 10: 45–63. https://doi.org/10.55630/dipp.2020.10.2.

Anders, Wallgren, and Wallgren Britt. 2007. *Register-Based Statistics. Administrative Data for Statistical Purposes*. 1st ed. Chichester:United Kingdom: John Wiley & Sons Ltd.

Antal, Daniel. 2020a. "Central And Eastern European Music Industry Report 2020." CEEMID, Consolidated Independent. https://doi.org/10.13140/RG.2.2.21450.31686.

———. 2020b. "Feasibility Study on Promoting Slovak Music in Slovakia & Abroad." https://doi.org/10.5281/zenodo.6427514.

———. 2023. "Pilot Program for Novel Music Industry Statistical Indicators in the Slovak Republic." Zenodo. https://doi.org/10.5281/zenodo.8399254.

Antal, Daniel, Mária Kmety Barteková, and Katarína Remeňová. 2023. "Economy of music in Europe: Novel data collection methods and indicators." Zenodo. https://doi.org/10.5281/zenodo.8334648.

Archives Expert Group on Archival Description, International Council on. 2023. "Records in Contexts–Conceptual Model. Version 1.0." International Council on Archives. https://www.ica.org/app/uploads/2023/12/RiC-CM-1.0.pdf.

Artisjus, HDS, SOZA, and Candole Partners. 2014. "Measuring and Reporting Regional Economic Value Added, National Income and Employment by the Music Industry in a Creative Industries Perspective. Memorandum of Understanding to Create a Regional Music Database to Support Professional National Reporting, Economic Valuation and a Regional Music Study."

Authority, International ISRC Registration. 2021. "International Standard Recording Code (ISRC) Handbook. 4th Edition." International ISRC Registration Authority. https://www.ifpi.org/wp-content/uploads/2021/02/ISRC_Handbook.pdf.

Bekiari, Chryssoula, George Bruseke, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, and Velios Athanasios, eds. 2024. "Definition of the CIDOC Conceptual Reference Model." CIDOC CRM Special Interest Group. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.4.pdf.

Camp, Ann Van, Sven Lieber, and IFLA. 2022. "ISNI, a Top Tool for Quality Enhancement, Smooth Data Flows and Efficient Internal Processes." Dublin: Ireland: International Federation of Library Associations; Institutions (IFLA). https://repository.ifla.org/handle/123456789/2008.

Cruz, Maria, and Clifford Tatum. 2021. "NWO Persistent Identifier Strategy." Zenodo. https://doi.org/10.5281/zenodo.4674513.

Curry, Edward. 2020. "Dataspaces: Fundamentals, Principles, and Techniques." In *Real-Time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems*, 45–62. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29665-0_3.

"Definition of the Europeana Data Model V5.2.8." 2017. Europeana. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf.

Diefenbach, Dennis, Max de Wilde, and Samantha Alipio. 2021. "Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph." In *ISWC 2021*. Online, France. https://hal.science/hal-03353225.

EBU, and Gaia-X. 2022. "Dataspace for Cultural and Creative Industries. Position Paper. v.2.0." Gaia-X. https://gaia-x.eu/wp-content/uploads/2022/10/EBU_position-paper_Media-Data-Space.pdf.

European Commission. 2021a. "Brochure for Music Moves Europe Preparatory Action 2019." European Commission. https://ec.europa.eu/culture/sites/default/files/library/mme_2019_brochure_final-web.pdf.

———. 2021b. "Music Moves Europe - First Dialogue Meeting. Final Report." European Commission. https://ec.europa.eu/culture/sites/default/files/library/mme-conference-report-web.pdf.

European Commission, Directorate-General for Communications Networks, Content, Technology, K Blind, S Pätsch, S Muto, M Böhm, T Schubert, P Grzegorzewska, and A Katz. 2021. *The Impact of Open Source Software and Hardware on Technological Independence, Competitiveness and Innovation in the EU Economy : Final Study Report*. Publications Office. https://doi.org/10.2759/430161.

European Commission, Directorate-General for Education, Youth, Sport and Culture, M Clarke, P Vroonhof, J Snijders, A Le Gall, B Jacquemet, et al. 2020. *Feasibility Study for the Establishment of a European Music Observatory : Final Report*. Publications Office of the European Union. https://doi.org/10.2766/9691.

Faraj, Ghazal, and András Micsik. 2023. "Enriching Wikidata with Cultural Heritage Data from the COURAGE Project." In, 407–18. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36599-8_37.

Fragkou, Pavlina. 2023. "DCAT-AP 3.0." Edited by Makx Dekkers, Pavlina Fragkou, Natasa Sofou, and Bert Van Nuffelen. https://semiceu.github.io/DCAT-AP/releases/3.0.0/.

Hartmann, Thomas, Sarven Capadisli, Franck Cotton, Richard Cyganiak, Arofan Gregory, Benedikt Kämpgen, Olof Olsson, Heiko Paulheim, Joachim Wackerow, and Benjamin Zapilko. 2024. "DDI-RDF Discovery Vocabulary. A Vocabulary for Publishing Metadata about Data Sets (Research and Survey Data) into the Web of Linked Data." Edited by Thomas Hartmann, Richard Cyganiak, Joachim Wackerow, and Benjamin Zapilko. W3C. https://rdf-vocabulary.ddialliance.org/discovery.html.

Huyer, Esther, and Laura van Knippenberg. 2020. *The Economic Impact of Open Data. Opportunities for Value Creation in Europe*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2830/63132.

ISO. 2017a. "ISO/IEC 19941:2017(en), Information Technology — Cloud Computing — Interoperability and Portability." London:United Kingdom: International Organization for Standardization. https://www.iso.org/obp/ui/#iso:std:iso-iec:19941:ed-1:v1:en.

———. 2017b. "ISO 2108:2017 (En), Information and Documentation — International Standard Book Number (ISBN)." International Organization for Standardization. https://www.iso.org/standard/65483.html.

———. 2019. “International Standard Recording Code (ISRC). ISO 3901:2019.” International Organization for Standardization. https://www.iso.org/standard/64817.html.

———. 2020. “ISO/IEC 22624:2020(en), Information Technology — Cloud Computing — Taxonomy Based Data Handling for Cloud Services.” London:United Kingdom: International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22624:ed-1:v1:en.

———. 2022. “International Standard Musical Work Code (ISWC). ISO 15707:2022.” International Organization for Standardization. https://www.iso.org/standard/83125.html.

———. 2023. “ISO/IEC 2382:2015(en), Information Technology — Vocabulary.” London:United Kingdom: International Standards Organisation. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:2382:ed-1:v2:en.

Kesäniemi, Joonas, Mikko Koho, and Eero Hyvönen. 2022. “Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC CRM.” In *New Trends in Database and Information Systems*, edited by Silvia Chiusano, Tania Cerquitelli, Robert Wrembel, Kjetil Nørvåg, Barbara Catania, Genoveva Vargas-Solar, and Ester Zumpano, 542–49. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-15743-1_49.

Ministerstvo kultúry SR, and Open Music Europe. 2023. “Memorandum o porozumení o využití výsledkov analýz otvorených politík v kontexte slovenského kultúrneho a kreatívneho priemyslu a sektorových verejných politík v spolupráci s konzorciom pre výskum a inovácie s názvom OpenMuse. [Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors’ public policies].” https://www.crz.gov.sk/zmluva/7645338/.

Nagel, Lars, and Douwe Lycklama, eds. 2021. “Design Principles for Data Spaces. Position Paper. Version 1.0.” Open DEI. https://doi.org/10.5281/zenodo.5244997.

Open Music Europe. 2023. “Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems.” https://doi.org/10.3030/101095295.

Pellegrino, Marco, and Denis Grofils. 2013. “DDI-SDMX Integration and Implementation. Working Paper.” United Nations Economic Commission for Europe. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP5.pdf.

Pomerantz, Jeffrey. 2015. “Definitions.” In *Metadata*, 19–64. Cambridge, MA, USA: The MIT Press. http://www.jstor.org.proxy.uba.uva.nl/stable/j.ctt1pv8904.6.

Rossenova, Lozana, Paul Duchesne, and Ina Blümel. 2022. “Wikidata and Wikibase as Complementary Research Data Management Services for Cultural Heritage Data.” In *CEUR Workshop Proceedings*. https://serwiss.bib.hs-hannover.de/frontdoor/deliver/index/docId/2573/file/rossenova_etal2022-wikidata_research_data_mgmt.pdf.

Siler, M. 2022. “Beyond the Fountain: Mapping a New Entry Point to the Society of Independent Artists.” *Art Documentation* 41 (2): 219–41. https://doi.org/10.1086/722172.

Sofou, Natasa, and Adina Dragan. 2019. “StatDCAT-AP – DCAT Application Profile for Description of Statistical Datasets. Version 1.0.1.” European Commission. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/statdcat-application-profile-data-portals-europe/release/101.

Stahl, Reinhold, and Patricia Staab. 2018. *Measuring the Data Universe: Data Integration Using Statistical Data and Metadata Exchange*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-76989-9.

UNECE. 2014. "Generic Statistical Information Model. GSIM V2.0 Documents. UNECE Statswiki." 2014. https://statswiki.unece.org/display/gsim/GSIM+v2.0+documents.

———. 2019. "Generic Statistical Business Process Model. GSBPM V5.2 Documents." UNECE Statswiki. January 2019. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. "Data Documentation Initiative: Toward a Standard for the Social Sciences." *International Journal of Digital Curation* 3 (1): 107–13.

Wickett, Karen M., Antoine Isaac, Katrina S. Fenlon, Martin Doerr, Carlo Meghini, Carole L. Palmer, and Jacob Jett. 2013. "Modeling Cultural Collections for Digital Aggregation and Exchange Environments." *CIRSS Technical Report 201310-1*, October. https://hdl.handle.net/2142/45860.

Wikimedia Foundation. n.d. "Wikibase Data Model." Wikimedia Foundation. Accessed May 26, 2024. https://www.mediawiki.org/wiki/Wikibase/DataModel.

ANNEXES

Annex 1 – Stakeholder profile data sheet for the Observatory Stakeholder Network

Name of the stakeholder:

- ☐ Corporate (institutional) name: This is mandatory for legal persons and groups; also, please provide a contact person.
- ☐ given name: mandatory for natural persons
- ☐ family name: mandatory for natural persons

Legal status of the stakeholder:

- ☒ private person: your name will be made public among the members; we also ask to provide a persistent ID (ISNI, ORCiD or VIAF.)
- ☒ Legal person: your name and legal person ID will be made public among the members, but not the contact person. Please provide ISNI, or OpenCorporates ID.
- ☐ other association or group without legal personality: your name and persistent ID will be made public among the members, but not that of the contact person. Please provide ISNI identifier.

Logo or icon of the stakeholder:

- ☐ not mandatory, but if provided, we will make it public

Online contact details of the stakeholder:

- ☐ Official website: we will make it public if provided
- ☐ LinkedIn page: we will make it public if provided
- ☐ Facebook page: we will make it public if provided
- ☐ YouTube channel: we will make it public if provided
- ☐ Instagram account: we will make it public if provided

Permanent residence or seat of the stakeholder:

- ☒ Only the country and municipality will be made public
- ☐ Please provide full postal address

Official email address of the stakeholder:

- ☐ only used for invitations to stakeholder meetings, giving and revoking data handling; never made public.

Contact telephone number of the stakeholder:

- ☐ is only used to clarify potential problems and consents; it is never made public and is not used unless necessary.

Annex 2 – Data Curators Manuals

Our main services involve data collection, processing, and dissemination. Without competent data curators, these services will not produce a high-quality data resource.

Data curators, as professionals, bear the responsibility of managing, maintaining, and enhancing the quality of an organisation's data. Their work is instrumental in making data easily accessible, accurate, and relevant to the organisation's needs. In large organisations, they collaborate closely with data engineers, data analysts, data scientists, and other stakeholders to establish a robust data ecosystem.



To work with the systems of the Open Music Observatory, we created two manuals.

- In March 2023 we created (and subsequently updated with user feedbacks) a Contributors Guideline for the internal stakeholders of the Open Music Europe (OpenMusE) project. This manual is available on contributors.dataobservatory.eu
- In 2024 we created a partly overlapping manual for the Observatory Stakeholder Network, i.e., for organisations that are providing and exchanging data with the Open Music Observatory. This manual is available on manual.opencollections.net.

In the music sector, because of the dominance of micro- and small enterprise (institution) sizes, very few competent data curators and specialised data or knowledge engineers are present. Our approach to solving this problem is the following:

- ☒ We pool those music experts within the stakeholder network who have data curatorial skills (for example, music librarians) or, due to their job, have background skills or an affinity to data curation.
- ☒ We provide these data curators with robust tools that only require a little learning.
- ☒ We centralise all the knowledge and data engineering work in the centre of the data-sharing space, i.e., at the Open Music Observatory.

We provide small-group training and online manuals for the data curators, responsible for maintaining the quality of data ingested by the Open Music Observatory. For large data providers, for example, collective management organisations or music information centres, we train one in-house data curator because, due to data confidentiality issues, often only an in-house person can review the totality of the data (and sort out the part of the data that can be shared within the dataspace with other observatory stakeholders.)

Inspiration

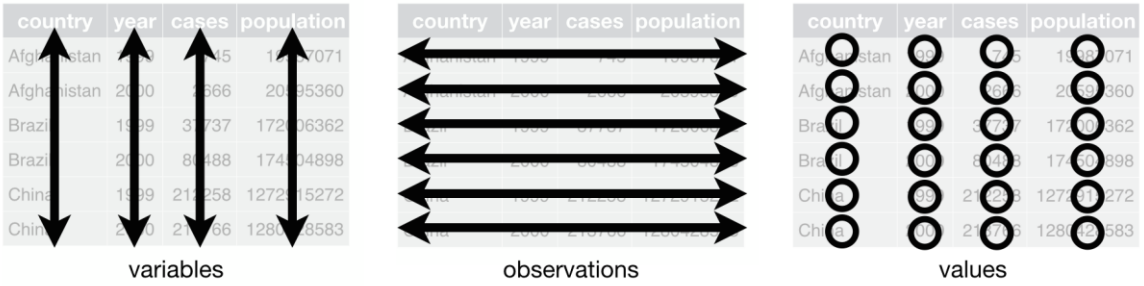
Very few music organisations employ data curators or employees with a library or information sciences background. We want to encourage music professionals working in various for-profit or social enterprises and research institutions to discover their "inner data curator." We believe that a passion for music and the sector and deep knowledge and experience with music are more important than the technical skills needed to curate the data.

We are encouraging the members of the *Observatory Stakeholder Network* (See: **?@sec-obstakeholdernetwork**) to find professionals, researchers, or artists in their organisation who have deep subject-domain knowledge about the data we want to improve: they know a lot about organs in

churches, about labels of a particular genre, sync licensing to films, or any other domain on which we collect data. Our ideal curators share a passion for data-driven evidence or visualisations, can learn tools that Wikipedia editors use, and have a robust and subjective idea about the data that would inform them in their work.

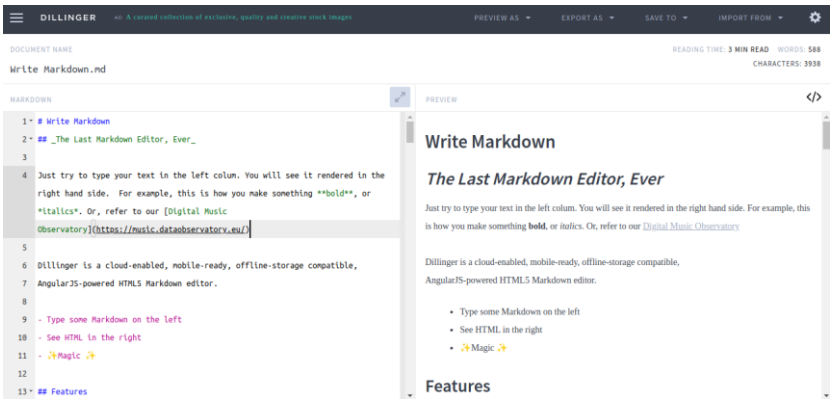
Basic Data Organisation Concepts

We are training and providing self-training material for two crucial but relatively simple concepts: tidy data and the text annotation and mark up.



Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells. From R For Data Science - 12. Tidy Data

Our documentation system works with MediaWiki, the mark up system developed by Wikipedia.

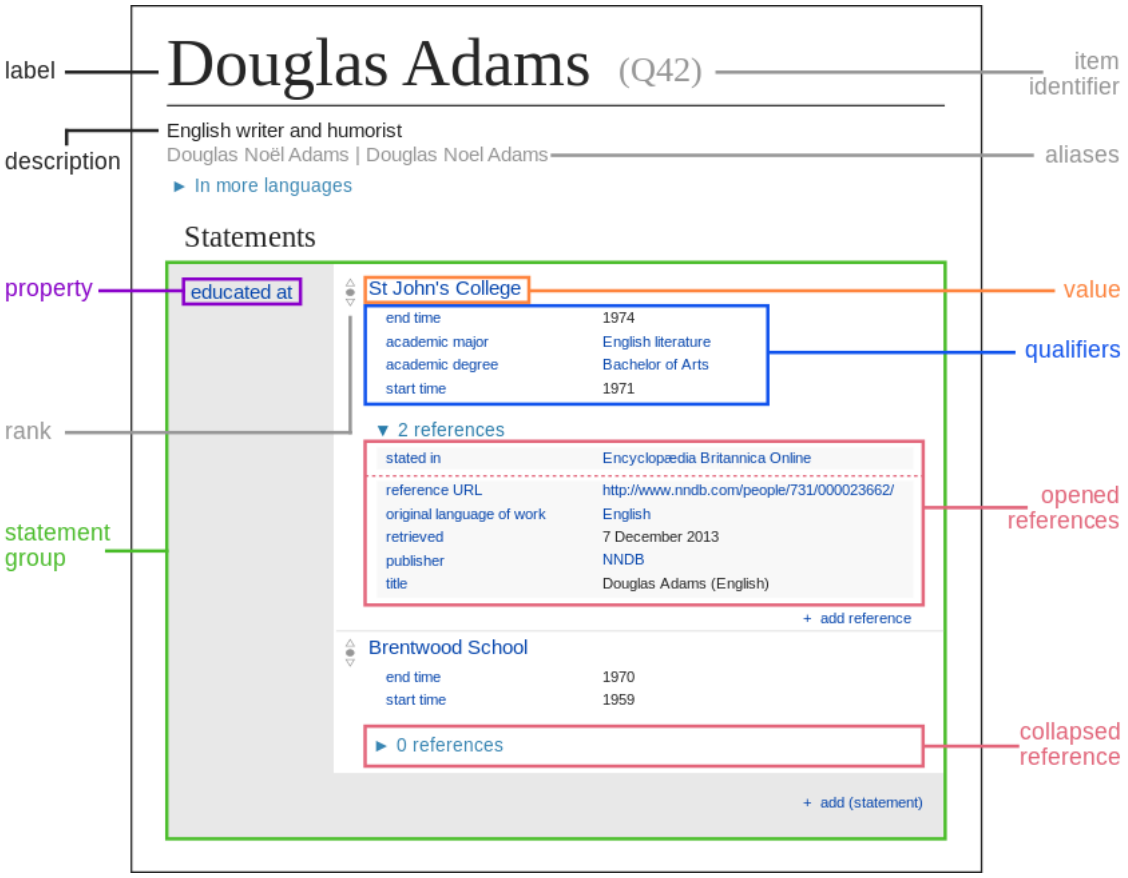


Dilinger is one of the best editors, and it is particularly suitable foor first-time markup users, as you immediately get visual feedback on how you mark up your text.

Understanding the Data Model

Wikidata is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation. It is a common source of open data that Wikimedia projects, such as Wikipedia, and anyone else, is able to use under the CC0 public domain license. As of early 2023, Wikidata had 1.54 billion item statements, or small, verifiable, scientific statements about our world. It runs on Wikibase, the tool that we use for the data consolidation of the Open Music Observatory.

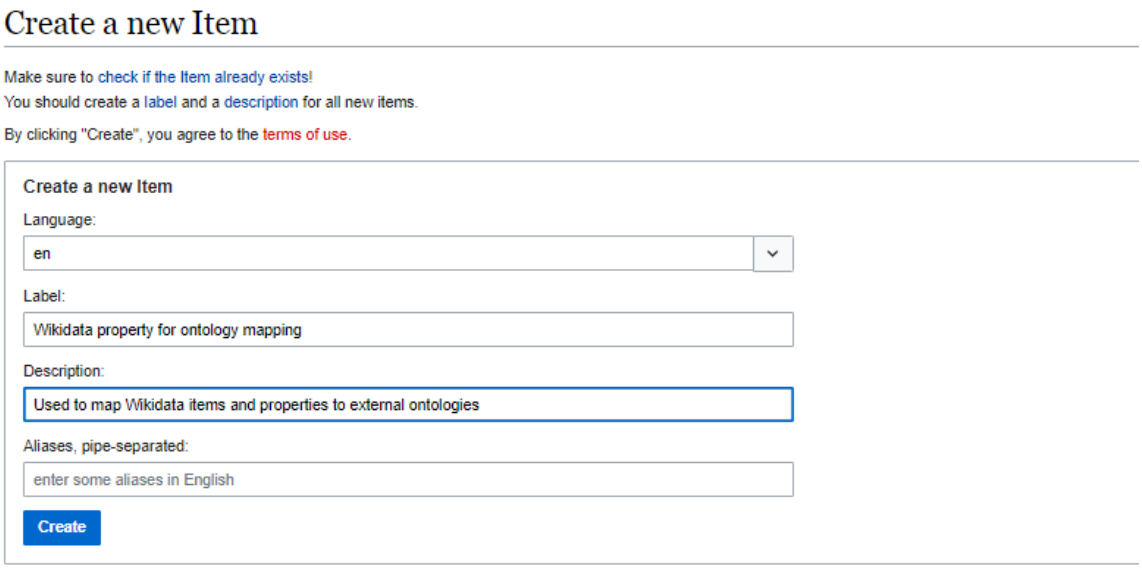
Wikidata is a document-oriented database, focusing on items, which represent any kind of topic, concept, or object.



Wikidata is a document-oriented database. This document connects a lot of knowledge about the late English writer and humorist, Douglas Adams.

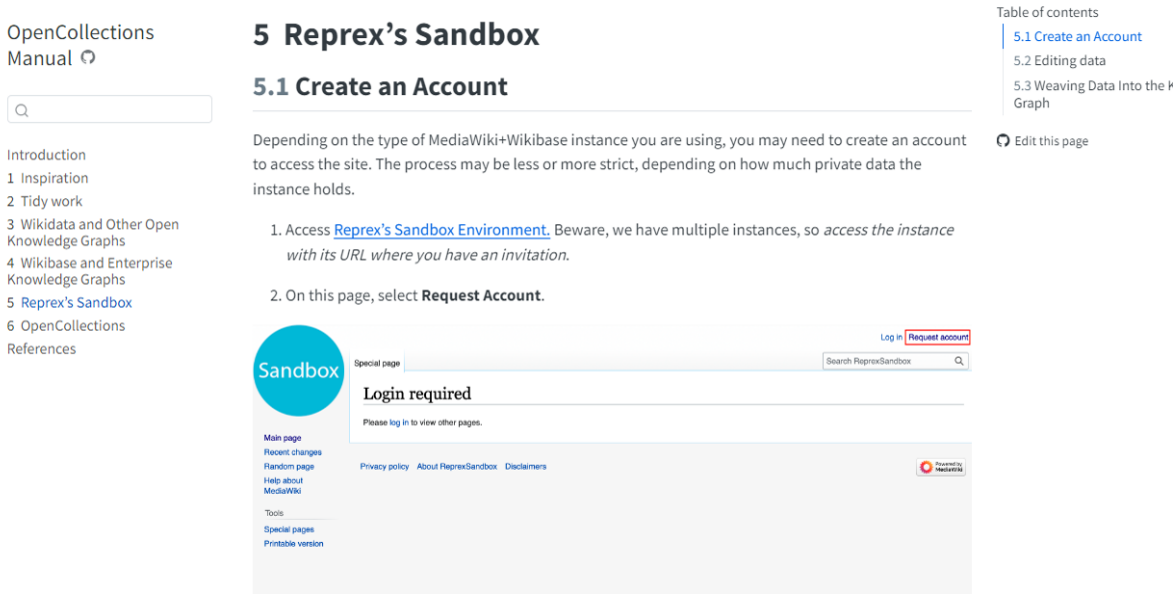
Data curators are expected to understand the basics of the Wikibase Data Model and the idea of working with a document-oriented database. We could learn from many EU and member state projects in this regard because Wikibase is a tool that is very often used for similar tasks. Originally intended at the level of citizen scientists, it allows music domain experts, like musicologists, music economists, music librarians and other non-technical stewards of data, to work efficiently with a data coordination system that uses Wikibase.

Working with the GUI



Identical to Wikidata: you must fill out at least the main Label of the item, and a description. We use English (en) as the master language for international cooperations.

Sandbox environment



Our manual is accompanied with a "sandbox" learning environment, where new data curators can try out various data manipuliation, editing, deleting, uploading actions without endangering their data, or the shared database.

Mass importing

Mass importing data requires solid technical skills because, in almost all cases, the data arrives from a very differently structured database: spreadsheets or a relational database management system. These tasks are performed by *Reprexbase* , the software components developed by Reprex to connect music

industry sources to the Wikibase Data Model. The guidelines provide information on how to prepare the data or what information should be given to us about the original schemas to start the mass import.

Data enrichment

The data enrichment are carried out with software components created by Reprex. It took about ten months to get enough data clearances in our data-sharing space to accumulate enough data for training enrichment algorithms. Results will be reported later in other tasks.

Quality Testing with SPARQL

SPARQL is the standard query language and protocol for Linked Open Data and RDF databases. Having been designed to query a great variety of data, it can efficiently extract information hidden in non-uniform data and store it in various formats and sources. SPARQL, pronounced 'sparkle', is the standard query language and protocol for Linked Open Data on the web or RDF triplestores. The SPARQL standard is designed and endorsed by the World Wide Web Consortium and helps users and developers focus on what they would like to know instead of how a database is organised.

Our data curators must be able to run SPARQL queries and make elementary modifications to them. Because we often import very large datasets, it would be very difficult to manually control every record on the graphical user interface. We use pre-written SPARQL queries (the data curator is expected to run via a simple URL link, perhaps modifying a class's QID or a property's PID) that serve as so-called *unit tests*. These queries programmed by Reprex allow simple tests like these:

- ☒ If the curator gave us 5432 person records, we have 5432 persons in the Reprexbase instance;
- ☒ If the gender breakup of a person's records is 2834:2598, the instance results in exactly the same persons of two genders (assuming that no third gender is used in the original data.)
- ☒ If we received data on Ján Levoslav Bella's Symphony in B minor, the publication year is 1982.

```
# Composers: citizens of Slovakia

SELECT ?item ?itemLabel  ?givenNameLabel ?lastnameLabel ?birthdate ?deathdate ?nationalityLabel
?itemDescription WHERE {
   ?item wdt:P31 wd:Q5 .            # instance of human
   ?item wdt:P106/wdt:P279* wd:Q36834.  # occupation or subclass of occupation that is composer
   ?item wdt:P27 wd:Q214.           # country of citizenship is Slovakia
   optional { ?item wdt:P735 ?lastname . }
   optional { ?item wdt:P734 ?givenName . }
   optional { ?item wdt:P569 ?birthdate . }
   optional { ?item wdt:P570 ?deathdate . }
   optional { ?item wdt:P27 ?nationality . }

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en,sk,de,hu" }
}

order by ?itemLabel
```

Try it out↗

Annex 3 – Terminology

Mapping guidelines

Music professionals

Wikidata uses the human (Q5) class a subclass of person (Q215627) that was defined much later. This often makes the mapping to CIDOC-CRM, RiC and many other ontologies ambiguous, because many persons are lacking a statement about their personhood. In collections management E21 Person (collections) and RiC-E08 (archives) are the most likely anchors of persons who are composers or performers of music.

For our use cases, the differentiation between living and deceased (not to mention imaginary) persons is necessary, so we encourage data curators to use the following mapping when importing their datasets:

- deceased person: use this preferred label to dead human (Q18093576). Our description: human who is no longer living (equivalent with dead person on Wikidata.)
- living person: use this preferred label to living human (Q18093573). Description: a human person who is alive, equivalent with living human on Wikidata.

We will add for both deceased person and living person the human statement for compatibility with Wikidata.

- imaginary person: use this preferred label to the imaginary character (Q115537581). Our description: character known only from narrations (fictional or in a factual manner) without a proof of existence; includes fictional, mythical, legendary or religious characters and similar; equivalent to the Wikidata item imaginary character. Imaginary persons are not entitled to copyright.
- deceased creator whose works are no longer protected by copyright: we create an inherited class of deceased person and creator. Creators who are living person or who do not belong to this class are assumed to have their works under copyright protection. Of course, in the case of multi-creator works, we cannot infer a public domain status from this class.

© 2024 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295