Open Music Observatory

Building an open data sharing space for the European music sector

Daniel Antal, CFA Mester, Anna Márta

2025-12-11

Table of contents

Open Music Observatory

Partly updated

This document was presented as a planning document in 2023. It has been partly updated till 18 November 2025. Some text is still reflecting the planning phase. You can access all versions on https://zenodo.org/records/16539570

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 Open Music Europe (OpenMusE) project partners as an account of work carried out within the framework of this contract.

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 Open Music Europe (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 Open Music Europe (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.

For the version history of this document, please refer to our open repository, where the change history can be reviewed with timestamps for every single file used to create the report: https://github.com/dataobservatory-eu/open-music-observatory

This document is accompanied by the maturing *A Green Paper on AI, Data Governance, and Metadata Policies for Europe's Music Ecosystem. Practical Steps Towards a Decentralised and Open European Music Observatory* document, which show's our works data-to-policy alignment. The latest development version of that document can be downloaded from the green paper's website, and the latest identified version following the OPA protocol can be found on GitHub and on Zenodo.

The temporary landing page of the OMO can be reviewed on https://dataobservatory-eu.github.io/omo-landing-page/.

Glossary

Music terms

`audio recording`: fixation of sounds (ISO 2019b)

`music video recording`: fixation of sounds synchronized with pictures or moving pictures where (a) the fixed sounds are wholly or substantially a musical performance or (b) the *recording* is intended for viewing in association with a recording of a musical performance. This definition includes music videos and concert recordings, together with music-related interviews and documentaries, but does not extend to genera! audiovisual material, even if it includes music.(ISO 2019b)

`recording`: result of a recording process independent of the type and number of audio or audiovisual carriers and technology used
Note 1 to entry: The term "recording" applies to each recorded item which may be used as a separate unit regardless of whether it is issued as part of a larger recorded work (e.g. each separate track on an album of audio recordings). [SOURCE:ISO 3901:2001, definition 3.3] (ISO 2017b)

`work`: distinct, abstract creation of the mind whose existence is revealed through one or more expressions (e.g. a performance) or manifestations (e.g. an object) (ISO 2022)

`musical work`: composed of a combination of sounds, with or without accompanying text (ISO 2022)

`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.

`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.

`duration`: the elapsed playing time between the first and last recorded modulations of the recording.

`LP` or `Long Player`: gramophone record usually on both sides comprising one or more sound recordings with a playing time of each side of normally round about 30 minutes and released and sold on its own (ISO 2017b)

`single`: gramophone record usually on both sides comprising one or two short sound recordings with a playing time of each side of normally no more than 7 minutes and released and sold on its own (ISO 2017b)

`track`: single recording on a sound carrier (ISO 2017b)

`original version`: The first established form of a work. (ISO 2022)

`medley`: A continuous and sequential combination of existing works or excerpts. (ISO 2022)

`potpourri`: A composite work with the addition of original material which have been combined to forma new work that has been published and printed. (ISO 2022)

`movement`: A principal division of a musical work. (ISO 2022)

`original title`: A title given to the work by its creator(s) shown in its original language. (ISO 2022)

`formal title`: A standardized title in which the elements are arranged in a pre-determined order, such as titles created for classical works. (ISO 2022)

`expression`: intellectual or artistic realisation of one and only one work
Note: may take the form of a notation , sound, image, object, movement or text (ISO 2017b)

`manifestation`: physical embodiment of an expression (ISO 2017b)

Creators of musical works

`arranger`: The author, or one of the authors, of an adapted text of a musical work. (ISO 2022)

`author`: The creator, or one of the creators, of the text of a musical work. (ISO 2022) Entitled to authors' right or copyright.

`composer`: The creator, or one of the creators, of the musical elements of a musical work. (ISO 2022) 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.

`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.

Data science terms

`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 formalized 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] (ISO 2019a)

`database`: collection of `data` organized 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] (ISO 2019a)

`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 definitions.* (ISO 2019a)

`datatype`: defined set of `data` objects of a specified data structure and a set of permissible operations, such that these data objects act as operands in the execution of any one of these operations (ISO 2019a);

data type is a class of data, characterized by the members of the class and the operations that can be applied to them (ISO 2017b)

`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.*

`data model`: description of the organization of data in a manner that reflects an information structure [SOURCE:ISO 28258:2013, definition 3.9] (ISO 2017b);

pattern of structuring data in a database according to the formal descriptions in its information system and according to the requirements of the database management system to be applied (ISO 2023b)

`data flow`: movement of data through the active parts of a data processing system in the course of the performance of specific work [SOURCE: ISO/IEC 2382:2015, 2121825] (ISO 2023b)

`big data`: extensive datasets – primarily in the data characteristics of `volume`, `variety`, `velocity`, and/or `variability`. – that require a scalable technology for efficient storage, manipulation, management, and analysis. note : Big data is commonly used in many different ways, for example as the name of the scalable technology used to handle big data extensive datasets. (ISO 2019a)

`data variability`: changes in transmission rate, format or structure, semantics, or quality of datasets (ISO 2019a)

`data variety`: range of formats, logical models, timescales, and semantics of a dataset. Note: Data veracity refers to descriptive data and self-inquiry about objects to support real-time decision-making. (ISO 2019a)

`data velocity`: rate of flow at which data is created, transmitted, stored, analysed or visualised (ISO 2019a)

`data volatility`: characteristic of data pertaining to the rate of change of these data over time [SOURCE:ISO/IEC 2382:2015, 17.06.06] (ISO 2019a)

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

`data science`: extraction of actionable knowledge from `data` through a process of discovery, or hypothesis and hypothesis testing (ISO 2019a)

`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] (ISO 2017a)

`cluster`: <distributed data processing> set of functional units under common control [SOURCE:ISO/IEC 2382:2015, 2120586] (ISO 2023b)

`scatter`: distribution of processing across multiple nodes in a cluster(ISO 2023b)

`file`: named set of records treated as a unit [SOURCE:ISO/IEC 2382:2015, 04.07.10] (ISO 2023b)

`document`: named, structured unit of text and possibly images that can be stored, edited, retrieved, and exchanged among systems or users as a separate unit (ISO 2023b)

`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 standardized by ISO/IEC [ISO/IEC 2382-1:1993]. (ISO 2023b)

`knowledge representation`: process or result of encoding and storing knowledge in a knowledge base. Term and definition standardized by ISO/IEC [ISO/IEC 2382-28:1995]. (ISO 2023b)

`knowledge graph`: a `knowledge representation` that uses a graph-structured data model to represent and operate on data. (ISO 2023b)

`knowledge source`: source of information from which a knowledge base has been created for a specific kind of problem Term and definition standardized by ISO/IEC [ISO/IEC 2382-28:1995]. (ISO 2023b)

`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 standardized by ISO/IEC [ISO/IEC 2382-28:1995]. (ISO 2023b)

`metadata`: data that define and describe other data (ISO 2023a); we use the more functional definition "a statement about a potentially informative object." (Pomerantz 2015b);

metadata is `data` about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and data volatility (ISO 2023b)

`structured data`: `data` which are organized based on a pre-defined (applicable) set of rules Note: The predefined set of rules governing the basis on which the data is structured needs to be clearly stated and made known. (ISO 2023b)

`partially structured data`: `data` that has some organization Note 1: Partially structured data is often referred to as semi-structured data by industry.
Note 2: examples of partially structured data are records with free text fields in addition to more structured fields. Such data is frequently represented in computer interpretable/parsible formats such as XML or JSON (ISO 2023b)

`horizontal scaling`: providing a single logical unit through the connection of multiple hardware and software. Note: The example of horizontal scaling is increasing the performance of distributed data processing through the addition of nodes in the cluster for additional resources. (ISO 2023b)

`vertical scaling`: act of increasing the performance of data processing through improvements to processors, memory, storage, or connectivity. (ISO 2023b)

`organization`: unique framework of authority within which a person or persons act, or are designated to act, towards some purpose [SOURCE:ISO/IEC 6523-1:1998, definition 3.1] (ISO 2017b)

`access level`: level of authority required from an entity to access a protected resource (ISO 2023b)

`algorithm`: finite ordered set of well-defined rules for the solution of a problem [SOURCE: ISO/IEC 2382-1:1993] (ISO 2023b)

`entity`: any concrete or abstract thing that exists, did exist, or might exist, including associations among these things. Example: A person, object, event, idea, process, etc.
Note 1 to entry: An entity exists whether data about it are available or not. [SOURCE:ISO/IEC 2382:2015, 2121433, modified] (ISO 2023a);

the entity is something capable of being uniquely identified
Note 1 to entry: Entities include material objects, electronic representations of content, abstract items (such as times, places), parties (human and corporate), as well as anything else that can be identified uniquely. (ISO 2017b)

`attribute`: characteristic of an object or set of objects [SOURCE:ISO/IEC 11179-3:2023, 3.1.11] (ISO 2023a)

`characteristic`: abstraction of a property. Example: 'Having a cable for connecting with a computer' as a characteristic of the concept 'cord mouse'. [SOURCE:ISO 1087:2019, 3.2.1] (ISO 2023a)

`property`: distinguishing feature of a material object (ISO 2017b)

`class`: description of a set of objects that share the same attributes, operations, methods, relationships and semantics (ISO 2023a)

`concept`: abstract entity for determining category membership
Note 1 to entry: A concept is used to classify objects (ISO 2023b);

a concept is a unit of knowledge created by a unique combination of characteristics (ISO 2017b)
Note 1 to entry: Concepts are not necessarily bound to particular natural languages. They are, however, influenced by the social or cultural background which often leads to different categorizations.

`relationship`: intellectual connection between two or more elements (ISO 2017b)

`relational model`: `data model` whose structure is based on a set of relations [SOURCE:ISO/IEC 2382:2015, 17.04.04] (ISO 2023b)

`non-relational model`: logical `data model` that does not follow a `relational model` for the storage and manipulation of data (ISO 2023b)

`identifier`: data string or pointer that establishes the identity of an item, organization or person alone or in combination with other elements [SOURCE:ISO 8459:2009, definition 2.27, modified] (ISO 2017b)

`persistent identifier` or `PID`: unique identifier that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership [SOURCE:ISO 24619:2011, definition 3.2.4] (ISO 2017b)

`qualifier`: information added to a descriptive element that assists its identification, understanding and/or use (ISO 2017b)

`taxonomy`: scheme of categories and subcategories that can be used to sort and otherwise organize itemized knowledge or information [SOURCE:ISO 25964-2:2013, definition 3.83 modified] (ISO 2017b)

`ontology`: formal, explicit specification of a shared conceptualization Note to entry: An ontology typically includes definitions of concepts and specified relationships between them, set out in a formal way so that a machine can use them for reasoning. [SOURCE:ISO 25964-2:2013, definition 3.57] (ISO 2017b)

`thesaurus`: controlled vocabulary and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.
Note 1 to entry: The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given

subject. For this reason a thesaurus is optimized for human navigability and terminological coverage of a domain. [SOURCE:ISO 25964-1:2011, definition 2.62] (ISO 2017b)

`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
Note 1: The term "expert system" is sometimes used synonymously with "knowledge-based system", but should be taken to emphasize expert knowledge.
Note 2. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems. (ISO 2023b) Expert systems fall under the definition of the AI Act.

`cloud computing`: paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand. (ISO 2019a)

`NERD`: named-entity recognition and disambiguation is a natural language processing technique that aims to resolve the ambiguity that arises from named entities in text.

Data protection terms

`DPIA`: Data Protection Impact Assessment (DPIA) is a process used to identify and minimize the risks associated with processing personal data.

`DPO`: the Data Protection Officer (DPO) is an individual designated by an organization to oversee its compliance with data protection laws, such as the GDPR. They act as a point of contact for data subjects and supervisory authorities, and they advise on and monitor data protection practices within the organization.

`GDPR`: The General Data Protection Regulation (GDPR) is a legal framework made by the European Union that sets guidelines for the collection and processing of personal information from individuals who live in and outside of the European Union.

Data curation and collection terms

`collection`: gathering of items assembled on the basis of some common characteristic, for some purpose, or as the result of some process (ISO 2017b)

`holdings`: totality of documents in the custody of an information and documentation organization (ISO 2017b)

`digital collection`: collection formed by a collection process on existing data and data sets where the collected data is in digital form (ISO 2017b)

`library collection`: all documents provided by a library for its users(ISO 2017b)

`anthology`: document consisting of a collection of full documents or of extracts, usually of literary works (ISO 2017b)

`exhibition`: curated display of objects on a clear concept and communicating a message [SOURCE:ISO 18461:2016, definition 2.4.6 modified] (ISO 2017b)

`curator`: person responsible for overseeing a collection or exhibition (ISO 2017b)

`data curation`: managed process, throughout the data lifecycle, by which data/data collections are cleansed, documented, standardized, formatted and interrelated (ISO 2017b)

`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;

a document, usually a volume, in which data are entered in a formal manner by a statutory authority Note 1 to entry: In modern usage, usually a database. (ISO 2017b)

`registration`: act of giving an entity a unique identifier on its entry into a system (ISO 2017b)

set of rules, operations, and procedures for inclusion of an item in a registry (ISO 2023a)

`registrant`: organization or person that has either registered an authentication protocol or registered the adoption of an authentication protocol [SOURCE: ISO/IEC 24727-6:2010, definition 3.4] (ISO 2017b);

an entity wishing to assign an ISRC to an applicable recording (ISO 2019b);

a `party` that requests an ISNI from the Registration Authority (ISNI 3.2 (ISO 2012, p15))

`party`: natural person or legal person, whether or not incorporated, or a group of either (ISO 2012)

`aggregation`: acquisition of sensitive information by collecting and correlating information of lesser sensitivity (ISO 2023b)

Statistical terms

`administrative records`: data generated by a non-statistical source, usually a public body, the main aim of which is not the provision of statistics.

`code list`: predefined list from which some statistical coded concepts take their values (ISO 2013)

`data pipeline`: a method in which raw data is ingested from various data sources and then ported to data store.

`FAIR` or `FAIR Guiding Principles for scientific data management and stewardship`: guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital

assets, emphasising machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention.)

`indicator`: the representation of statistical data for a specified time, place or any other relevant characteristic, corrected for at least one dimension (usually size) so as to allow for meaningful comparison.

`microdata`: non-aggregated observations or measurements of characteristics of individual units, without direct identifier.

`MVP` or `minimum viable product`: a version of a work product with just enough features and requirements to satisfy early customers and/or provide feedback for future development [SOURCE:IEEE 2675-2021, 3.1]

`observation unit`: an identifiable entity about which data can be obtained, it is also often called a *statistical unit* or *data subject* in case of a natural person.

`Open Policy Analysis Guidelines`: a set of information management rules to make policy analysis more transparent.

`personal data`: any information relating to an identified or identifiable natural person.

`pseudonymisation`: processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information.

`survey`: a systematic examination and record of a physical or social area and its features so as to construct a map, plan, or description. In social sciences it usually refers to a well-structured questionnaire and answers given to its items by a target population.

`statistics`: quantitative and qualitative, aggregated and representative information characterising a collective phenomenon in a considered population.

`visualisations`: schematic charts, drawings, photographs, and their collages will as still image files that help to explain the relationship between information carriers, data points, or processes.

Registers, authorities, standards and identifiers

`IČO`: The organisation identification number (IČO) is an identifier assigned to all types of legal entities, entrepreneurs and public authorities by the Statistical Office of the Slovak Republic. The Czech Republic's organisation identifier is also called IČO.

`OpenCorporates`: a public corporation database which sources data from national business registries.

`ISNI`: an ISO certified global standard number for identifying the millions of contributors to creative works and those active in their distribution.

`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.

`VIAF ID`: The VIAF (Virtual International Authority File) combines multiple name authority files into a single OCLC-hosted name authority service.

`ISRC`: The International Standard Recording Code (ISRC) is a standard identifying code that can be used to identify sound recordings and music video recordings so that each such recording can be referred to uniquely and unambiguously.

`ISWC`: The purpose in creating an ISWC for musical works is to enable more efficient administration of rights to those works on a worldwide basis. The ISWC provides an efficient means of identifying musical works in computer databases and related documentation and for the exchange of information between rights societies, publishers, record companies and other interested parties on an international level.

`ISBN`: the International Standard Book Number is an identification system for the publishing industry and its supply chains.

`ISMN`: The International standard music number (ISMN) was developed by, and for, the music publishing sector as a separate system to complement the International standard book number (ISBN). The existence of the ISMN as a separate identifier system makes it possible to identify printed and notated music as a distinct category of publication within the global supply chain and to develop trade directories and similar services for the specialized market for music publications.

`ISCC`: The International Standard Content Code (ISCC) is an identifier for numerous types of digital assets.

`DOI`: The Digital Object Identifier is a standardised unique number given to many (but not all) articles, papers and books, by some publishers, to identify a particular publication.

`ORCID`: the Open Researcher and Contributor ID is a unique, persistent identifier free of charge to researchers.

`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.

`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.

`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.

`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.

`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.

`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.

`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 EU.

`CIDOC-CRM`: The conceptual model of CIDOC, the standard conceptualisation of collection management systems in heritage organisations.

`RiC`: *Records in Context* is a new conceptual model that replaces the four most important international archiving standards.

`DCTERMS` or `DCMI`: the Dublin Core Metadata Terms is a vocabulary of metadata terms developed and maintained by the Dublin Core Metadata Initiative (DCMI). These terms are used to describe various aspects of digital resources, such as web pages, documents, and other online content. They provide a standardized way to assign metadata to resources, making them easier to discover, manage, and exchange.

`RDFS`: the Resource Description Framework Schema is an extension of the Resource Description Framework (RDF) that provides a vocabulary for describing classes and properties of resources within an RDF graph.

`EDM`: the Europeana Data Model is a framework for collecting, connecting, and enriching cultural heritage metadata. It's designed to facilitate the sharing and reuse of cultural heritage information by providing a standardized way to represent and link data.

`Europeana`: a digital platform provided by the European Union that aggregates digitized cultural heritage from institutions across Europe.

`ESCO`: the European Skills, Competences, Qualifications and Occupations classification is is a multilingual classification system developed by the European Commission to standardize the description of skills, competences, and qualifications relevant to the European labor market and education.

NACE: the European Union's standard classification of economic activities for statistical purposes. The abbreviation stands for Nomenclature statistique des Activités économiques dans la Communauté européenne.

ISCO: the International Standard Classification of Occupations (ISCO) is the International Labour Organization's standardized system for classifying and organizing occupations according to jobs' tasks and duties

ISIC: the International Standard Industrial Classification of All Economic Activities (ISIC) is a standard classification system developed by the UN Statistics Division (UNSD) to categorize economic activities.

PROV-O: the Provenance ontology is a formal ontology developed by W3C to represent and interchange provenance information.

MARC: MAchine-Readable Cataloging, is a standard digital format used by libraries to represent and exchange bibliographic information.

DCAT: an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.

Organisations

AEPO-ARTIS: Organisation representing European artists-performers. Regroups most of the European CMO representing performers.

ALOADED: is a company which distributes and exploits recordings.

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.

CNM (former CNV): the Centre National de la Musique is a public organisation managing a tax on concert tickets

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 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 analysing 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.

GESAC: GESAC comprises together 32 European authors' societies in music, audiovisual, visual arts, literature and drama.

GESIS: Leibniz Institute for the Social Sciences.

IAML: International Association of Music Libraries, Archives and Documentation Centres |

IAMIC: International Association of Music Centres, an international network of organisations that collectively and collaboratively provides information and promotes the music of their countries or regions.

ICMP: the global trade body representing the music publishing industry worldwide.

SCAPR: International association for the development of the practical cooperation between performers' collective management organisations (CMOs)

SOZA: SOZA (Slovenský ochranný zväz autorský pre práva k hudobným dielam, Slovak Performing and Mechanical Rights Society) is a legal entity, non-profit civic association of authors and publishers of musical works, association of natural persons and legal entities.

Hudobné Centrum: Music Centre Slovakia is a music organisation with a mission to promote Slovak contemporaly music.

Other abbreviations

CEEMID: the Central European Music Industry Databases is a multi-country project that was a predecessor of Reprex's Digital Music Observatory

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.

DSP: Digital service providers (DSPs), or Digital Streaming Platforms are companies or organisations that provide access to services online.

EIF: The European Interoperability Framework (EIF) is a set of recommendations and guidelines that aims to facilitate communication and collaboration between public administrations, businesses, and citizens within the European Union and across national borders.

ECCCH: The European Collaborative Cloud for Cultural Heritage is a European Union initiative for a digital infrastructure that will connect cultural heritage institutions and professionals across the EU.

EOSC: The European Open Science Cloud (EOSC) aims to create a trusted, open, and multidisciplinary environment for researchers and innovators in Europe.

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.

RDM: Research Data Management refers to the suite of practices, policies, and processes used to handle data throughout the lifecycle of a research project.

Our glossary is harmonised with relevant music-sector specific standards (referred to in Chapter 8) and with the ISO *Information technology — Vocabulary* (ISO 2023b); *Information technology — Cloud computing — Taxonomy based data handling for cloud services* (ISO 2020); *Information technology — Cloud computing — Interoperability and portability* (ISO 2017a) and the *Information and documentation — Foundation and vocabulary* (ISO 2017b) and Information technology — Metadata registries (MDR) — 1. Framework (ISO 2023a)

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.

In the OpenMusE project, the development of the Open Music Observatory is coupled with a clear contractual expectation: the project must populate the Observatory's four thematic pillars—*Music Economy, Music Diversity, Music, Society & Sustainability, and Innovation & Future Trends* —with initial, well-documented data and knowledge. This population process follows the project's data-to-policy pipeline (see Chapter 1): each work package defines its indicators, establishes data governance and legal bases, collects or accesses relevant administrative, survey, statistical, and platform data, processes and harmonises them through WP4 tools, and finally activates them in reproducible analytical workflows in WP5. The result is that the Observatory is not only a technical prototype but a functional, data-bearing infrastructure: the first integrated demonstration of how Europe's music data can be curated, linked, analysed, and made reusable across public, private, and civic actors.

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.

The *European Interoperability Framework* aims to create a four-layered approach to build digital research, marketing, rights management, collection management services for the sector. These layers are introduced in separate chapters of these documentation.

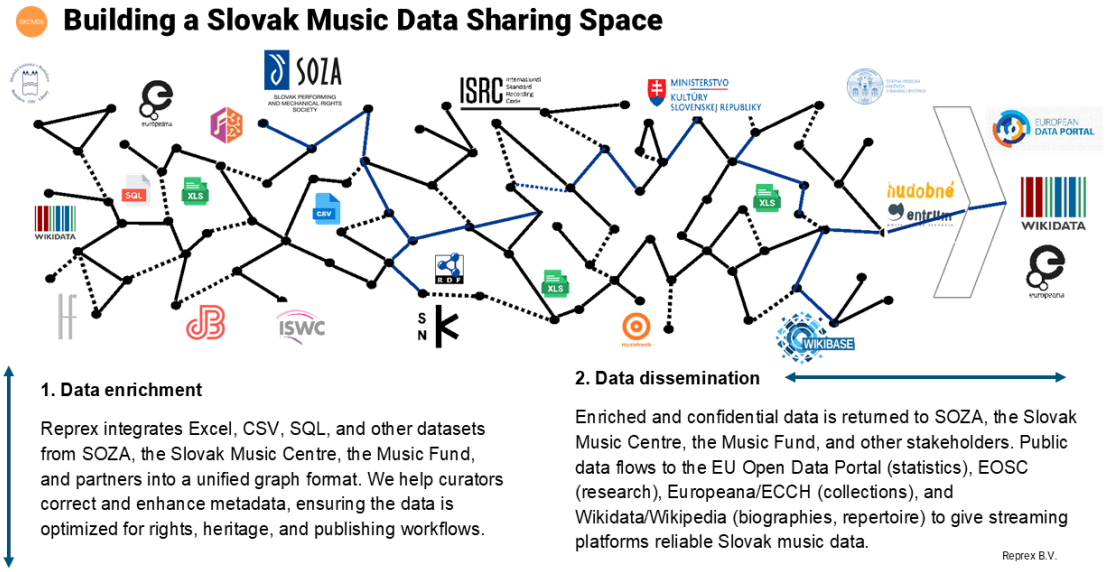
1. The **technological alignment** is introduced in Chapter 4; we decided to choose the technology of the world's largest open knowledge graph, Wikidata, which already coordinates countless digital services in Europe's cultural sector, and provides training and guardrails for many AI applications.
2. The **semantic alignment** is is introduced in Chapter 5. This chapter focuses on the semantic and organisational aspects of interoperability.
3. The **organisational** alignment means bridging actual data-driven and computer supported workflows with data semantics (what does a song's *title* mean for a librarian, an ethnomusicologist, a collective rights management agency), and how they can work with various translated, alternative, historical, mistyped, and preferred titles in distributing royalties, loaning printed sheets, describing musical traditions.
4. The **legal** alignment creates a policy that lays out the rights, prohibitions and necessary permission processes to connect and use the data together. We give a concrete example in Section 5.3.

We were informed and influenced by the creation of *Europeana* (which started out from a similar collaborative project, a cultural heritage oriented data sharing space) and the Commission's new plans to extend their digital services into the *European Collaborative Cloud for Cultural Heritage* (ECCCH).

We aimed for full interoperability with *Europeana* and we Reprex successfully sent data and concluded a Data Exchange Agreement. We were also aiming for interoperability with the ECCCH, which only published the first version of its Heritage Digital Twin conceptual model and ontology; we were the first to test them with music data.

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.

By the end of 2025, we aim to create an "observatory 3.0", which already hosts many intelligent data improvement technologies and fuels innovative applications/services in line with our project's exploitation pathways. 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.



1 Introduction

The *Open Music Observatory* (OMO) is the technical and organisational backbone of the Open Music Europe (OpenMusE) project. It is conceived as a federated, standards-based data space that integrates the project's work on data governance, reproducible data collection, semantic harmonisation, and policy translation. Rather than a single centralised database, the Observatory serves as a *shared interoperability layer* where data from administrative, statistical, survey, heritage, and platform sources can be linked through persistent identifiers, common vocabularies, and transparent provenance. Its purpose is to provide the practical foundations for a future European Music Observatory by enabling decentralised actors to contribute data in a trustworthy and reproducible way.

The development of the Open Music Observatory has also converged strongly with the findings of the **CITF First Project Report**, even though the two initiatives worked in parallel and address different sectors. The CITF report defines the requirements for trustworthy, machine-readable copyright infrastructures in the AI era, especially the need for reliable rights metadata, lifecycle-based provenance, persistent identifiers, and clear links between public-sector collections and private registries. The Observatory implements these same principles within the music domain: it establishes federated governance across institutions, creates verifiable reference objects for works, recordings, and contributors, aligns rights and heritage metadata, and embeds provenance throughout all ingestion and harmonisation workflows. In this sense, the Observatory provides a practical, domain-specific implementation of several CITF requirements, demonstrating how a decentralised data space can support copyright, AI governance, and cultural-heritage interoperability at the same time. This alignment is intentional and positions the Observatory as a complementary, sector-specific counterpart to the cross-sector framework developed by the CITF, and it is explained in detail in Chapter 2.

The Observatory is populated through a multi-stage workflow that follows the project's data-to-policy pipeline (see Section 2.5). First, data curation work (WP6 and WP1–WP4) establishes the Data Management Plan, licensing conditions, data protection rules, and methodological standards. This is followed by the development of reproducible data-collection and processing tools (WP3 and WP4), including survey workflows, register harmonisation, streaming-data pipelines, and automated ingestion scripts. Once datasets are curated, validated, and harmonised, they are integrated into the Observatory's semantic layer and linked to indicators used in policy analysis (WP5). The slide reproduced in this report illustrates this sequence clearly: (1) **data curation** → (2) **methodology and indicator design** → (3–4) **reproducible collection and processing** → (5–7) **OMO design and policy analysis** → (8) **dissemination**.

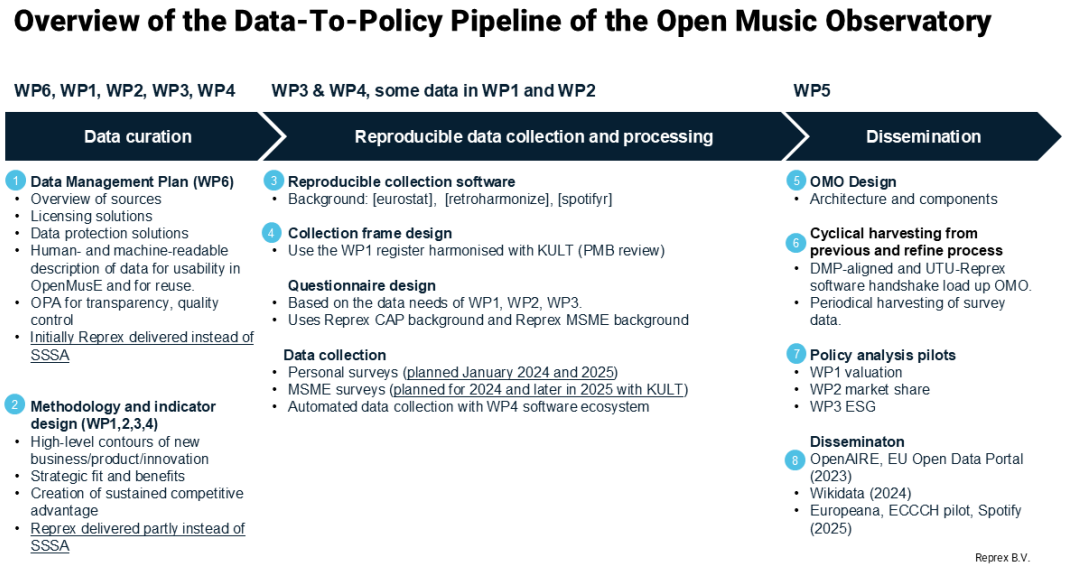


Figure 1.1: The data-to-policy pipeline of the OpenMusE project

This technical report documents how this process works in practice.

The Background chapter (Chapter 2) explains the conceptual and policy foundations for a decentralised observatory and traces the evolution from CEEMID to the Open Music Observatory.

The **Architecture** chapter (Chapter 4) describes how the observatory is technically designed—why Wikibase is used, how semantic interoperability is achieved, and how automated ingestion supports the pipeline.

The **Data Catalogue** (Chapter 10) presents the datasets already curated and ready for ingestion, while the Data Sources chapter (**?@sec-datasources**) explains how administrative, statistical, survey, and platform data were identified and how they align with the Data Management Plan.

The chapters on **Data Collection and Processing** and on **Standardisation** describe the software tools delivered in WP4 and the harmonisation schemas that make interoperability possible. Finally, the Dissemination and Dataspace chapters present how the Observatory integrates with Europeana, the EU Open Data Portal, EOSC, ECCCH, and other infrastructures, as well as how national pilots—such as the Slovak Comprehensive Music Database—demonstrate how the Observatory can scale.

Together, these chapters show that the *Open Music Observatory* is not a static repository but a *living, federated infrastructure*: one that connects data, metadata, workflows, and policy outputs, and one that can be sustainably expanded by institutions across Europe. This report provides the full technical, methodological, and policy context needed to understand

how the Observatory is being built and how it can serve as the basis for a future European Music Observatory.

Following the *Executive summary* and this introduction (Chapter 1), the report is structured as follows:

Chapter 2 provides an overview of the historical plans for establishing a European music observatory and the rationale for a decentralized approach. It introduces CEEMID—an important precursor to the Open Music Europe project and a key building block identified in the *EMO feasibility study*. We then conceptualize the Open Music Observatory as a data-sharing space, a technical and legal innovation of the European Union. This section emphasizes compliance with the European Interoperability Framework and the adoption of open collaboration methods, open-source software, and open data as fundamental components. Finally, we review a selection of existing observatories, including those assessed in the *EMO feasibility study*, to compare their current services with our own.

Chapter 5 explains why the development of the *Open Music Observatory* requires not only technical tools but also data coordination across institutions, and why such coordination must follow a federated rather than centralised model for music and cultural data. It outlines how distributed governance, subsidiarity, and institutional autonomy shape data practices in Europe, making coordination essential for reliability, comparability, and equitable representation. The chapter introduces the *European Interoperability Framework* (EIF) as the conceptual backbone for aligning legal, organisational, semantic, and technical layers across diverse actors, and shows how these principles can be extended beyond the public sector to include collective rights organisations, private distributors, memory institutions, statistical offices, and community actors. We believe that a *data sharing space* is the the practical mechanism for enabling this alignment—distributed infrastructures where actors retain stewardship over their data but interoperate through shared identifiers, semantic mediation, and provenance-aware workflows.

Chapter 3 offers an overview of the current core services of the Observatory, which function as those of a modern statistical vendor and open data portal. In line with the *European Interoperability Framework* (EIF), we integrate our services into the EU Open Data Portal (as the primary gateway for disseminating statistical datasets), Europeana and the European Cultural Heritage Cloud (for music collections), and the Wikipedia ecosystem (Wikidata and Wikibase) to distribute data into open knowledge graphs.

Chapter 10 introduces our digital curation guidelines and the nucleus of our data offering. Before the release of Observatory 2.0, we could not ingest large, high-value datasets, so the current catalogue contains datasets derived solely from the background IP of the Open Music Europe project. These datasets are already available for testing, and we plan to present the first large-scale datasets and onboarding case studies at several international conferences in November 2025. One such large-scale case study is already underway and will be introduced later in Section 3.1.1.

?@sec-datasources revisits the *EMO feasibility study* and provides an updated overview of data availability and potential partners across the EU. At the time of the feasibility study,

many pan-European music organizations pledged to contribute data to a future observatory. We will present the new Open Music Observatory to them, invite them to join the Observatory Stakeholder Network, and request at least sample datasets for publication.

Chapter 8 outlines the data and terminology standardization work initiated by the Open Music Observatory. The feasibility study correctly identified this as a key missing function. We divide the standardization effort into three main areas:

- ☒ Standardization of music-related conceptual and information models for the data we collect and disseminate;
- ☒ Standardization of entity identities, including author names, works, and recordings;
- ☒ Standardization of music-related and cross-domain concepts, such as "average income," "composer," or "lyricist."

Chapter 9 introduces the "Observatory 3.0" concept, presenting our planned value-added services designed to deliver trustworthy data and leverage trustworthy AI for the benefit of the music sector. We also highlight innovative services that consortium partners are developing using the machine-actionable data our APIs will provide.

The *Slovak Comprehensive Music Database* `SKCMDb` https://hudobnadatabaza.sk/ is a national initiative to make Slovak music more accessible, discoverable, and usable across libraries, archives, streaming platforms, and rights organisations. As a functional component of the Observatory, it demonstrates how the dataspace model can integrate public, private, civic, and community data under shared governance. A detailed description is provided in the (see in the Annex).

The *Livonian Music Database* (`LīvMDb`) shows that our solution can work wiht music that has low documentation depth, weak institutions. The music of the Livonian people is scattered, and as native speakers of this small ethnic group died out, their heritage was dispersed, and largely not placed on modern digital platforms. The `LīvMDb` as a functional module of the *Open Music Observatory*, serves as a testbed for working with very low documentation, decolonisation, and other issues related to regional cultures, ethnic minorities. Its experiences can directly be translated to work with any low documentation depth environment, for example, catalogues of microlabels, self-releasing artists, small publishers. It offers insights into subsidiarity (see in the Annex).

The annexes contain supporting material on the implementation status of the pipeline, the relationships between services and standards, and the governance of the Observatory.

Together, this report and Deliverable 5.1 demonstrate how the Open Music Observatory implements the Grant Agreement's vision and how it prepares the foundation for a European Music Observatory.

2 Background & Concept

This chapter provides the historical, conceptual, and policy foundations needed to understand why and how the Open Music Europe project operates through a unified data-to-policy pipeline, and why the *Open Music Observatory* (OMO) and the *Open Music Europe Data Management Plan* (DMP) form the backbone of that pipeline.

It first explains the origins of the European Music Observatory idea, tracing the policy dialogue from *Music Moves Europe* and the 2018 *EMO Feasibility Study* to the decentralised CEEMID initiative, which anticipated many of the technical and governance principles later adopted by the EU's data-space strategy. It then situates *Open Music Europe* within the evolving European policy environment: the European Parliament's resolutions, the EU Music Ecosystem Study, the emergence of sectoral data spaces, and alignment with Europeana, EOSC/ECCCH, and the EU Open Data Portal.

This context is essential for understanding why the Observatory must be *federated, decentralised, interoperable, and automated*, and why a traditional centralised observatory model is no longer viable.

The temporary landing page of the OMO can be reviewed on https://dataobservatory-eu.github.io/omo-landing-page/.

Policy Context: The Open Music Europe Green Paper

The *Green Paper on AI, Data Governance, and Metadata Policies for Europe's Music Ecosystem* provides the broader policy framework for the Open Music Observatory.
It explains:

- how the data-to-policy pipeline aligns with the EU's *European Strategy for Data*,
- how OMO's federated architecture reflects the principles of the *Data Governance Act*, *Data Act*, and the *Data Spaces Support Centre (DSSC)*,
- how metadata governance interacts with *AI policies* (AI Act, Trustworthy AI Guidelines, cultural-sector risk considerations), and
- how OMO supports emerging reforms in *copyright infrastructures*, especially where attribution, identifiers, and lifecycle metadata intersect.

The Green Paper also integrates insights from the *CITF First Project Report (2025)*, showing how music-sector data spaces complement cross-sector requirements for trustworthy, machine-readable rights metadata.
Together with the OMO technical documentation and the Data Management Plan, the Green Paper forms the *policy and governance layer* of the Observatory's design.

Finally, the chapter introduces the *Open Music Europe data-to-policy pipeline*—the core methodological mechanism defined in the Grant Agreement as "an open, scalable data-to-policy pipeline for European music ecosystems." The pipeline structures all project activities (WP1–WP5) and provides the operational logic of the Observatory. It links:

- indicator design,
- data governance,
- data acquisition,
- metadata and semantic modelling,
- statistical analysis, and
- policy translation,

all under the governance framework established by the DMP.

Taken together, these elements explain the rationale behind the OMO architecture, its dataspace foundations, and the methodological choices documented in the remainder of the report.

2.1 Why Europe Needs a Music Observatory

In late 2015, the European Commission began a structured dialogue with representatives of the music sector[1] to identify key challenges and possible forms of EU support. Under the *Music Moves Europe* programme, the Commission launched the 2018 Preparatory Action *Boosting European Music Diversity and Talent*, which produced the *Feasibility Study for the Establishment of a European Music Observatory* (European Commission et al. 2020). The later *CITF First Project Report (2025)* independently identified similar structural obstacles across copyright-dependent sectors: fragmented identifiers, inconsistencies in rights metadata, and the absence of lifecycle-aware registries. This convergence strengthens the evidence base for an Observatory design rooted in federation, interoperability, and transparent provenance.

The Feasibility Study identified **45 data gaps** that impede the development of evidence-based policies for competitiveness, value creation, diversity, and employment in the European music ecosystem. It highlighted that:

> "Data collection in Eastern and Southern Europe is lagging in comparison to other European Member States in Northern and Western Europe… 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." (European Commission et al. 2020, p9–10)

[1] See (European Commission 2021b, 2021a) on the policy-making process and objectives; *Feasibility Study for the Establishment of a European Music Observatory* (European Commission et al. 2020) and the *Interoperable, Trustworthy, and Machine-Readable Copyright Data in the AI Era: Report of the CITF First Project* (Partanen et al. 2025).

The study explicitly mentioned CEEMID as a promising bottom-up model for filling these gaps and providing a more modern, decentralised alternative to traditional observatory structures (Artisjus et al. 2014).

The *Feasibility Study* also provided a clear definition of the stakeholders that a future *European Music Observatory* must serve. It distinguished three groups whose information needs and policy roles must be supported:

- **Industry**: commercial organisations and agents involved in income-generating activities across performance, recording, distribution, and creation.
- **Civic**: policymakers, NGOs, professional associations, and publicly funded intermediaries whose decisions shape the regulatory and support environment.
- **Public**: consumers, cultural participants, education and training institutions, and third-sector organisations interested in the wider social and cultural roles of music.

See (European Commission et al. 2020, p30)

These stakeholder categories continue to structure the Observatory's service model and interoperability requirements.

2.2 Historical Precedent: CEEMID

The former **CEEMID** (originally: *Central & Eastern European Music Industry Databases*) collaboration began in 2014 as a voluntary, decentralised data initiative created by three collective management societies. Over time, it grew to include more than 60 stakeholders in 12 European countries. Its purpose was to fill the most pressing evidence gaps by combining:

- voluntary data integration among partners,
- open-data reprocessing, and
- co-financed data collection.

This work is documented in (Antal 2020a).

CEEMID operated according to principles that would later become central to the European Union's **data (sharing) space** strategy (formalised only years later). Its decentralised organisational model, distributed data stewardship, and emphasis on transparent, reusable methods demonstrated that a modern observatory in the digital era does not need to be a centralised institution. Instead, it can function as a federated ecosystem connecting statistical offices, cultural institutions, CMOs, and private actors.

Long before the EU formalised its dataspace strategy, CEEMID also aligned its workflows with emerging statistical-system standards such as **GSIM**, **DDI**, and **SDMX**, anticipating later European requirements for interoperable, machine-readable statistical metadata. This early adoption provided a methodological bridge between cultural-sector data, administrative registers, and official statistics, and formed a direct precursor to the metadata foundations of the *Open Music Observatory*.

The *Feasibility Study for the European Music Observatory* explicitly recognised CEEMID as a potential building block for a new observatory model. Our proposal therefore sought to transform CEEMID's prototype—referred to in the study as the *Digital Music Observatory*—into a scientifically robust and methodologically coherent system that could scale across Europe. This required grounding the work in state-of-the-art statistical science, data science, and computer science, and ensuring alignment with European interoperability and data-governance frameworks.

The prototype work that preceded *Open Music Europe* was shaped through two innovation environments: the *Yes!Delft AI+Blockchain Lab*, where product–market fit and technical feasibility were tested, and the *JUMP Music Market Accelerator*, where the first integrated prototype of a *Digital Music Observatory* was developed. These early iterations validated not only stakeholder demand but also the feasibility of a decentralised, standards-based architecture, and they informed the methodological and technical design choices taken forward in this project.

2.3 Policy and Technological Evolution Enabling a New Observatory

Since the publication of the *EMO Feasibility Study*, the European Union has introduced a series of policy and infrastructure initiatives that strengthen the case for a **decentralised, interoperable, and federated** European Music Observatory. These developments span European Parliament mandates, Commission-funded research, cultural-heritage clouds, open-data regulation, and the EU's overarching data-space strategy. Together, they establish the policy and technological foundations on which the *Open Music Observatory* is built.

Our policy alignment is discussed in more detail in - *Music Metadata Mainstreaming and EU Law - A Green Paper on AI, Data Governance, and Metadata–Policies for Europe's Music Ecosystem*[2]

2.3.1 European Parliament and EU-level Mandates

The **European Parliament**, in its resolutions on the future of the music sector, explicitly called for:

- the establishment of a *European Music Observatory*,
- improved evidence for competitiveness, diversity, and fair remuneration, and
- stronger coordination of public, private, and community data sources.

These mandates update and reinforce both the *Music Moves Europe* framework and the findings of the *EMO Feasibility Study*. They frame the Observatory as an instrument that must serve industry, civic, and public actors through interoperable, reusable, cross-border data services.

[2]See (Senftleben et al. 2024); and (Antal 2025f), summarised in the internal document (Open Music Europe Consortium 2025).

The *EU Music Ecosystem Study* (2025) deepened this diagnosis, pointing to fragmentation across metadata, rights information, cultural statistics, and market data. It concluded that the sector requires a technical and governance model capable of linking these domains, rather than separate, siloed initiatives. The architecture of our dataspace responds directly to these recommendations.

The European Parliament has rightly highlighted that fragmented and unreliable metadata remains a major obstacle in the music sector. *European Parliament Resolution of 17 January 2024 on Cultural Diversity and the Conditions for Authors in the European Music Streaming Market*

> 9. Emphasises that it is essential to improve the identification of anyone involved in the creation process, in particular authors and performers, on music streaming services, by ensuring the comprehensive and accurate allocation of metadata from the time of Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multiterritorial licensing of rights in musical works for online use in the internal market (OJ L 84, 20.3.2014, p. 72). creation for any track uploaded to a music streaming service; encourages, in this regard, the use of all international identification codes (IPI, ISWC, ISRC, IPN, and ISNI); highlights that proper identification of creators plays a key role in the search for and discoverability of works, and enables proper remuneration for creators in the distribution of revenues.

Our Observatory's distributed model directly answers European Parliament's call for metadata systems that are reliable, inclusive, and supportive of creators. Our policy alignment is explained in detail in our our policy paper, *A Green Paper on AI, Data Governance, and Metadata Policies for Europe's Music Ecosystem*[3].

2.3.2 Data (Sharing) Spaces

The EU's adoption of **data (sharing) spaces** provides the organisational and legal model for an Observatory that is not a centralised institution but a **federated ecosystem**.

Curry defines dataspaces as:

> "an emerging approach to data management… Data is integrated on an 'as-needed' basis, with the labour-intensive aspects of data integration postponed until they are required." (Curry 2020)

The *Design Principles for Data Spaces* position paper further describes them as:

> "a federated data ecosystem within a certain application domain and based on shared policies and rules." (Nagel and Lycklama 2021, p7)

[3]The Music ecosytem study: (Music Moves Europe 2024); the European Parliament's resolution (European Parliament 2024) and our policy paper: (Antal 2025f).

These principles are fully consistent with CEEMID's decentralised model and form the conceptual basis for the `Open Music Dataspace` (see Chapter 5).

The *CITF (2025) report* arrives at the same architectural conclusion. It stresses that trustworthy copyright infrastructures in the AI era require federated governance, interoperable identifiers, and verifiable provenance chains rather than a single, centralised registry. Its three-layer model—foundational identifiers, shared semantics, and technical services—maps closely onto the Observatory's dataspace design.

Observatories created in the 1990s and early 2000s were built around centralised databases and slow-moving data-collection cycles. Since then, the rapid expansion of **agentic AI in data collection**, the widespread digitisation of live and recorded music, and the proliferation of **large-scale, real-time data sources** have made such centralised architectures obsolete. Modern evidence ecosystems require automated ingestion, continuous semantic enrichment, cross-domain reconciliation, and transparent provenance — all of which presuppose a federated, decentralised model rather than a single institutional database.

The European Audiovisual Observatory (EAO), the European Market Observatory for Fisheries and Aquaculture Products (EUMOFA), and the European Observatory on Infringements of Intellectual Property Rights (EUIPO) provide valuable models of long-standing EU observatories. However, each operates within a centralised data-submission and aggregation framework appropriate to their legal mandates and sectoral data structures. The Feasibility Study acknowledged that the music sector lacks comparable legal obligations and contains far more fragmented, cross-domain, multilingual, and institutionally diverse datasets. Therefore, while these observatories offer important governance precedents, their centralised architectures cannot be replicated in the music ecosystem — strengthening the case for a federated dataspace model.

2.3.3 Preference for Open-Source and Open Standards in the EU

Across the EU's data and digital-transition strategies, there is a consistent preference for:

- open-source software,
- open standards,
- open licensing, and
- transparent, reproducible workflows.

This aligns directly with the Observatory's use of open-source R and Python pipelines, Wikibase for semantic interoperability, and FAIR-compliant metadata.

2.3.4 Alignment with Europeana and Cultural Heritage Infrastructures

Europeana demonstrates how Europe manages distributed cultural-haritage collections at scale using:

- persistent identifiers,
- multilingual metadata,

- open licences (e.g. CC BY),
- shared semantic standards (EDM, IIIF, rightsstatements.org), and
- decentralised stewardship by libraries, archives, and museums.

The *Open Music Observatory* follows the same principles. It uses:

- semantic technologies,
- PID-based cross-domain linking, and
- open, reusable data models.

This ensures interoperability with cultural-heritage collections, performing-arts archives, and national memory institutions, and aligns the music domain with the emerging **European Collaborative Cloud for Cultural Heritage (ECCCH)**.

2.3.5 Alignment with the EU Open Data Portal and EU Open Data Strategy

The **EU Open Data Portal** (data.europa.eu) establishes a common framework for:

- open licences (e.g. CC BY 4.0),
- machine-readable formats,
- harmonised metadata (DCAT-AP),
- and publication of public-sector information.

The *Open Music Observatory* is designed so that:

- public datasets can be harvested directly by the EU Open Data Portal,
- indicators and derived datasets comply with open-data rules, and
- metadata follow **DCAT-AP** and **DataCite** to support long-term reuse.

This alignment ensures that the Observatory meets both Horizon Europe open-science requirements and broader EU open-data policy objectives.

2.3.6 European Interoperability Framework (EIF)

The **European Interoperability Framework (EIF)** provides a four-layer model—legal, organisational, semantic, technical—for connecting:

- public authorities,
- cultural institutions,
- rights-management organisations,
- national statistical offices, and
- private intermediaries.

These are precisely the actors whose data must interoperate to support a European Music Observatory. By adopting the EIF, the Observatory can link diverse datasets into coherent, reusable services without centralising them.

2.3.7 EOSC and ECCCH: Open Science and Cultural-Heritage Clouds

The **European Open Science Cloud (EOSC)** and the **European Collaborative Cloud for Cultural Heritage (ECCCH)** promote:

- FAIR data,
- open science workflows,
- reproducible analysis,
- transparent provenance, and
- decentralised storage and processing.

These principles inform the Observatory's architecture through the use of:

- open-source analytical pipelines,
- SDMX and DataCite metadata,
- persistent identifiers, and
- federated linking across domains and institutions.

2.3.8 Summary

Together, these EU policy instruments—the Parliament's mandate, the EU Music Ecosystem Study, data-space strategy, Europeana, the EU Open Data Portal, the EIF, EOSC, and ECCCH—provide a unified rationale for an Observatory that is federated, decentralised, data-driven, and interoperable by design. They define the policy and technological environment in which the *Open Music Observatory* must operate and directly shape its architecture. The Chapter 4 explains why we chose an architecture that is built around Wikibase and Wikiadta.

2.4 Why Open Music Europe Uses a Decentralised Dataspace Model

The *Open Music Observatory* adopts a decentralised, federated dataspace model because this is the only architecture that meets the needs identified by the EMO Feasibility Study, the EU Music Ecosystem Study, and the European Parliament's resolutions, while also complying with the newer EU frameworks for interoperability, data governance, and cultural-heritage infrastructures. A centralised database model, common in observatories built in the 1990s or early 2000s, is no longer feasible or desirable for the music sector.

2.4.1 Lessons from CEEMID

The CEEMID collaboration demonstrated that most music-sector data—repertoire, rights, cultural-heritage descriptions, business metadata, and statistical evidence—originate from many different institutions, each with its own mandates, legal obligations, and technical systems. Centralising such data is:

- legally constrained (e.g. GDPR, contractual confidentiality),
- institutionally unrealistic (distributed ownership and stewardship), and
- technically inefficient (rapidly evolving local systems).

CEEMID showed that these data can nonetheless be made interoperable through:

- shared identifiers and authority files,
- open metadata standards,
- reproducible R-based pipelines, and
- rule-based, voluntary data sharing.

These are the foundational principles of a **data (sharing) space**, which the EU has since elevated to a core strategic component of its digital-policy agenda.

2.4.2 Requirements of the EU policy environment

As outlined in Section C, the EU now expects cultural and creative sectors to adopt:

- federated data architectures,
- FAIR and open data practices,
- transparent governance models,
- semantic interoperability, and
- alignment with Europeana, EOSC, ECCCH, and data.europa.eu.

This expectation reflects the broader transformation of European data governance, where sectors are encouraged to organise around **data spaces** rather than central repositories. A decentralised model also supports cultural and data sovereignty by allowing institutions to maintain control over their collections and data-processing rules.

2.4.3 Requirements of the Grant Agreement

The *Open Music Europe* Grant Agreement defines the project explicitly as:

> "an open, scalable data-to-policy pipeline for European music ecosystems"

and mandates the creation of:

> "a highly automated, decentralised intelligence hub that aggregates open data and creates dynamic, live policy documents."

To fulfil these contractual obligations, the Observatory must:

- connect heterogeneous data sources without centralising them,
- refresh indicators automatically as upstream data changes,
- maintain legally sound provenance across many institutions,
- support multilingual, cross-border metadata, and
- integrate statistical, cultural-heritage, and industry systems.

These requirements can only be met in a **federated dataspace**, not in a single, centralised database.

2.4.4 Technical rationale for decentralisation

The dataspace model makes it possible to:

- keep **sensitive or personal data** (e.g. rights, royalties) within the institution that controls them,
- link sources through **semantic federation** (Wikibase/Wikidata),
- enable distributed curation by librarians, archivists, CMOs, and researchers,
- integrate permanent identifier (PID) systems across domains (ISNI, VIAF, ROR, company registers),
- use open standards (SDMX, DDI, DataCite, DCAT-AP), and
- scale to new partners, genres, languages, and Member States.

This structure mirrors the actual distribution of data in the music sector and the technical direction of the EU's digital transition.

Recent research highlights how music discovery is increasingly shaped by opaque, platform-controlled recommendation systems that structure visibility, attention, and cultural participation. These systems exert measurable influence on user behaviour, commercial outcomes, and the availability of minority or non-mainstream repertoires, yet they remain largely inaccessible to independent scrutiny. The argument that public-interest infrastructures must provide transparent, auditable, and diversity-preserving alternatives aligns directly with the rationale for a decentralised European Music Observatory. By foregrounding metadata quality, open workflow documentation, and federated governance, the Observatory responds to concerns that current algorithmic environments reproduce structural asymmetries and limit cultural plurality (Guest, Suarez, and Rooij 2025).

2.4.5 Why decentralisation is essential for a European Music Observatory

For the European music ecosystem, decentralisation enables:

- lower administrative and compliance burdens,
- institutional autonomy and data sovereignty,
- cross-border comparability without forced data transfer,
- community- and expert-driven metadata improvement,
- GDPR-compliant handling of personal data, and
- sustainable expansion of the Observatory.

A decentralised dataspace is therefore not an architectural choice but a **necessary governance model** for an Observatory that spans cultural heritage, rights management, statistical registers, community archives, and private-sector metadata across the EU.

The *Open Music Observatory* is consequently designed as a **federated, rule-based dataspace**: an ecosystem where public, private, and civic stakeholders contribute knowledge, maintain authority records, and generate indicators while preserving full control over their own data.

2.5 The Data-to-Policy Pipeline: How Open Music Europe Works

The *Open Music Europe* action is contractually defined as **"an open, scalable data-to-policy pipeline for European music ecosystems"** (see Grant Agreement). This is not a slogan: it is the methodological core of the project and the organising principle of all work packages (WP1–WP5). The pipeline connects indicator design, data governance, data acquisition, semantic modelling, statistical analysis, and policy translation into a single reproducible workflow. This chapter introduces the logic of that pipeline and explains how it shapes the design of the *Open Music Observatory*.

2.5.1 1. Indicator and problem definition (WP1–WP3)

Each thematic work package begins by identifying policy-relevant gaps and defining the indicators needed to address them. Deliverables D1.1, D2.1, and D3.1 specify:

- the conceptual frameworks guiding each domain (economy, diversity, society),
- the data requirements for measuring them, and
- the procedures for ensuring comparability across countries and years.

These definitions also appear in the *Open Music Europe Data Management Plan* (D6.3), which provides human-readable summaries and machine-readable metadata for all indicators.

2.5.2 2. Data governance (WP1–WP3, WP6)

Before data can be collected or integrated, partners agree on:

- sources, access rights, and sampling frames;
- metadata standards (SDMX, DDI, DataCite);
- ethical safeguards and GDPR-compliant procedures;
- controlled vocabularies, authority files, and persistent identifiers.

These agreements are formalised in D1.2, D2.2, D3.2, and the Data Management Plan (D6.3). They ensure compliance with FAIR, OPA, and EU data-governance principles.

2.5.3 3. Software for data collection (WP4)

WP4 develops the open-source tools used to gather and ingest administrative data, survey data, platform usage data, and CMO records. These tools form the operational backbone of the pipeline. They include:

- survey-data management scripts,
- connectors for royalty and licensing accounts,
- streaming API integration modules,
- metadata templates for ingestion and harmonisation.

All tools adhere to the reproducibility and interoperability requirements defined in Annex 1 and the DMP.

2.5.4 4. Data acquisition (WP1, WP2, WP3)

Data are collected from:

- collective management organisations (CMOs),
- ministries and statistical offices,
- cultural-heritage institutions,
- surveys (enterprise and personal),
- streaming-service APIs.

Each domain follows its own protocol (e.g. WP1 T1.2 sampling frames; WP3 T3.1 participation and wellbeing indicators). Data collected are documented in the DMP and cross-referenced with OPA-compliant folders.

2.5.5 5. Processing, enrichment, and harmonisation (WP4, WP5)

Raw inputs are processed using REPREX's R-based *openmusic-pipeline*:

- cleaning and pseudonymisation,
- metadata harmonisation,
- cross-linking with authority records,
- structuring in SDMX/DataCite formats,
- integration via persistent identifiers.

This step transforms heterogeneous inputs into consistent, analysis-ready datasets.

2.5.6 6. Validation for analysis and dissemination (WP4, WP5)

Before modelling can begin, WP4 and WP5 validate:

- interoperability across sources,
- semantic consistency,
- statistical reproducibility,
- versioning and provenance,
- integration with the *Open Music Observatory*.

This ensures that indicators can be reproduced from source data and that all transformations are transparent.

2.5.7 7. Analysis and modelling (WP1–WP3)

The three thematic work packages conduct analyses based on their indicator sets:

- WP1: music-sector economic valuation, satellite accounting extensions, export and employment metrics.
- WP2: diversity, circulation, repertoire mapping, multilinguality.
- WP3: participation, wellbeing, social impact, SDG alignment.

These analyses produce the evidence base for policy recommendations and are documented in open, executable notebooks.

2.5.8 8. Policy translation (WP5)

WP5 converts analytical results into **Open Policy Analysis (OPA)** documents — open, versioned, executable notebooks linking:

- data,
- code,
- statistical results,
- interpretative text.

Deliverables D5.6 and D5.7 demonstrate this approach. These "living policy documents" integrate directly with the *Open Music Observatory* and update as data refresh.

2.5.9 9. Dissemination and reuse (WP5)

Final datasets, indicators, and notebooks are published:

- through the *Open Music Observatory*,
- on Zenodo with DOIs,
- and, where applicable, on the EU Open Data Portal.

This ensures transparency, auditability, and long-term reuse by policymakers, researchers, and cultural-sector stakeholders.

2.5.10 Summary

Taken together, these nine stages describe the way *Open Music Europe* operates: a **unified, reproducible pipeline** that transforms fragmented, multi-source data into validated, policy-ready evidence. The pipeline is the methodological backbone of the project and the operational logic of the *Open Music Observatory*.

2.6 Stakeholder Engagement and Early Feedback

Throughout the development of the *Open Music Observatory*, the consortium sought feedback from representative stakeholders across the music ecosystem, cultural-heritage institutions, data-infrastructure initiatives, and the open-knowledge community. Key presentations included:

- **Networkshop 2024**, Eger, Hungary — possibilities of a cross-border federation between Slovak and Hungarian music data spaces (Antal 2024a).
- **IAMIC General Assembly and Conference 2024**, Vienna — introduction of the Slovak Comprehensive Music Database and early Observatory concepts (Antal 2024c).
- **IAML Annual Conference 2025**, Salzburg — presentation and poster on interoperability of music libraries and archives with public and private music services (Antal 2025a, 2025b).
- **Polifonia Stakeholder Session (2023)** — metadata and interoperability workshop with performing-arts archives, national memory institutions, and academic partners.
- **Big Data Value Association / Gaia-X / Data Space Support Centre: Data Week 2024**, Leuven — introduction of a European music dataspace and alignment with DSSC principles.
- **CISAC European Committee Meeting 2025**, Vilnius — discussion of metadata governance and creators' rights in the AI era (Mikš 2025).
- **Wikimedia CEE Meeting 2024**, Istanbul, and **Wikimedia CEE Meeting 2025**, Thessaloniki — technical presentations to Wikidata and Abstract Wikipedia communities (Antal 2024b).
- **Wikidata Conf 2025** (online),showing how our data model and governance can improve Wikidata itself.
  These consultations helped validate the Observatory's technical, semantic, and governance foundations and informed its alignment with sectoral needs.

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.

The role of data interoperability is not to exchange or centralise data, but to create digital services in libraries, rights management, distribution, streaming platforms that work together well. For the service providers they should work with less data processing cost, and for the user they should give a higher value or better user experience

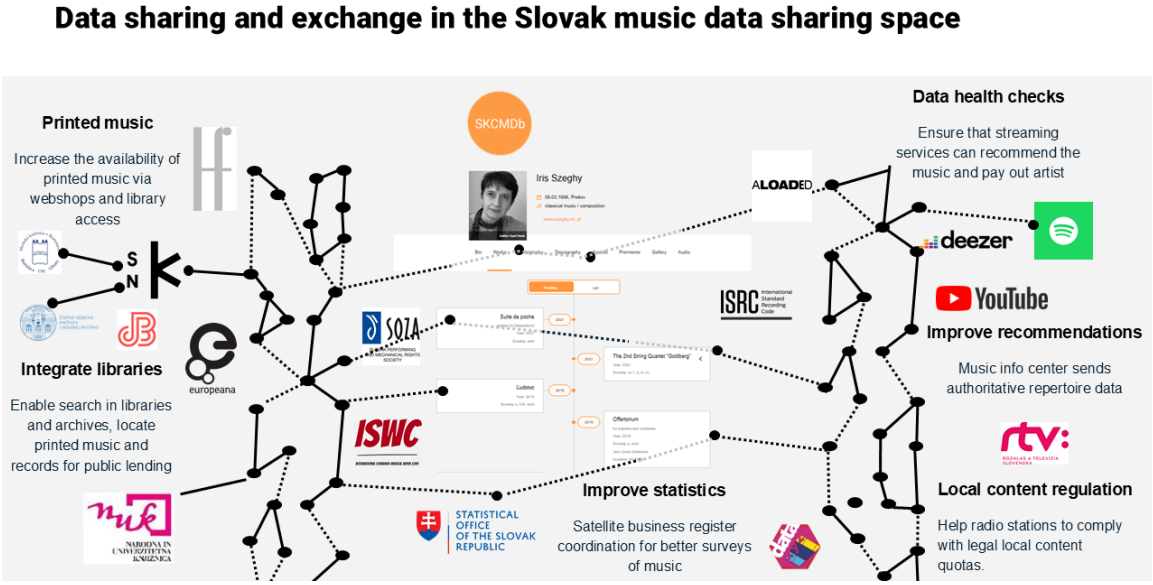


Figure 3.1: The role of data interoperability is not to exchange or centralise data, but to create interoperable digital services (illustration repeated from earlier section.)

The *Open Music Observatory* aims to improve data interoperability so that it can help music researchers, policymakers, and industry players to:

- Reduce the cost of maintaining public or private services, including knowledge management and reproducible research, and even evidence-based policy-making.

- Increase the value or user experience of public or private services with a higher diversity of music available, with more evidence and knowledge present for further research and business or public policy evidence.

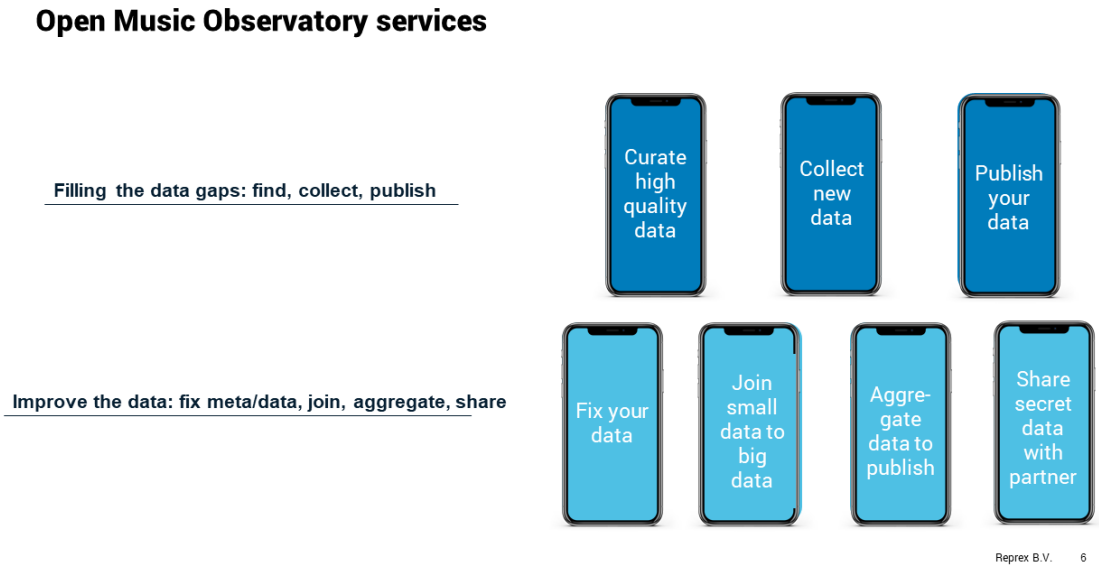
An important way that can help to serve both needs is the harmonisation of costly registration processes and services (Antal and Mester 2025).

> ⚠ Not updated
>
> This section was created at an early planning stage, and had not yet been updated. This is no longer applicable, and should not be read or quoted. However, However, this part is clearly articulated in our green paper's Fixing Music Data at the Source and the AI that Works for Music, Not Against It chapters (Antal 2025f).

The *Slovak Comprehensive Music Database* (`SKCMDb`) is a national initiative to make Slovak music more accessible, discoverable, and usable across libraries, archives, streaming services, and rights organizations. It connects scores, recordings, and metadata using open standards and collaborative governance. As a functional module of the Open Music Observatory, SKCMDb also serves as a testbed for developing shared services.

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.

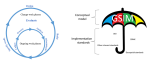


Figure 3.2: The two implementation standards, DDI and SDMX ensures that the Open Music Observatory can work together with statistical offices, or respectable social sciences data repositories.

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 9.

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.

> **Note**
>
> 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.

> **Note**
>
> 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 10.1. 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 8.

3.1.1 Microdata, Collections, Records

We treat "microdata" as a collection of structured data.

A **register** is a document [in modern usage, usually a database], in which data are entered in a formal manner by a statutory authority (ISO 2017b). In statistical data collection ian 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.

A **collection** is a group of objects, for example, musical works, sound recordings, printed scores, 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; music labels and publisher select and musical works and their recordings or scores to place into commercial circulation. Such collections form the basis of census or sample surveys for statistical data collection.

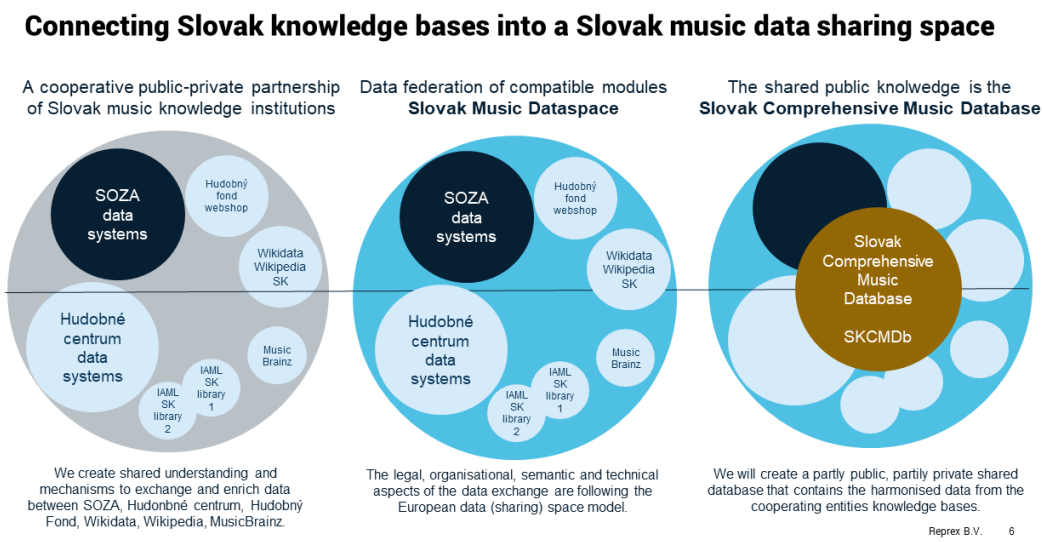
The documentation of **collections** relies on the work of **registers**. For example, music publishers can claim their revenues based on ISWC and ISMN identifiers provided to them by the collective management organisations that register works, or national libraries or other organisations that identify printed sheets.

The maintenance of registers requires ongoing investment, and therefore registrars like the ISRC Authority or CISAC (the manager of the ISWC register) often restrict access to their data, or do not exchange data. In an increasingly globalised, automated music ecosystem where the number of identifiable works, recordings, scores, and related claims is growing exponentially, this situation puts the entire industry at a disadvantage, for example, against tech platforms. The Open Music Observatory is experimenting with innovative ways how registers can work together in some aspects of metadata standardisation, improvement and exchange in a way that keeps their core product intact and exclusive to them See: (Antal and Mester 2025).

The Open Music Observatory works with metadata in a way that helps managing and improving registers, and it helps to create data about collections with authoritative data from registers.

> **Note**
>
> Our first large database 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 is partly based on registers, and partly on curated holdings of Slovak stakeholders.



- ☒ Our collections are always available on https://reprexbase.eu/skcmdb/ . Further details in Section 3.5.1.
- ☒ 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 Collaborative Cloud for Cultural Heritage.

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 8.

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 2015a, 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 2015a, 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.

> **i** Note
>
> **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

> ⚠ Warning
>
> 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 Repair

Throughout the project we realised that data and metadata repair is perhaps a more urgent challenge then data processing.

While our team and our stakeholders gradually embraced the concept of a data sharing space, i.e., the idea that instead of starting new data collections from scratch it is more economic and useful utilise existing data, given the high level of music industry digitalisation and that almost all transactions leave a digital trail behind, we also realised that the music sector is "drowning in numbers"; it handles more data in various obsolete, undocumented, unstructured or ad hoc forms than it can utilise. In these cases, usually the data is already available somewhere, but in a format that prevents the data to be used to its potential.

Most of our efforts therefore were concentrated on data and metadata repair instead of new collection. Metadata repair and data processing is usually hard-to-distinguish tasks that comprise of similar or same steps. They are various validation, normalisation procedures that allow that make the data informative.

Recent analyses show that gaps, inconsistencies, and ambiguities in underlying metadata can propagate directly into algorithmic systems, affecting how repertoires are ranked, surfaced, or made visible to listeners. Situating repair as a core service within the Observatory therefore responds not only to documentation needs but also to wider concerns about equitable discovery conditions in platform-mediated music environments (Guest, Suarez, and Rooij 2025).

3.3 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.

3.3.1 Processing & re-processing microdata

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

The EU Open Data Portal uses the following namespace definitions; these definitions refer to machine readable, explicit definitions (ontologies) of the way our datasets must be understood by a software agent. To demistify the process, we provide here an example of the metadata that we need to compile from the various steps of the data production pipeline.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix adms: <http://www.w3.org/ns/adms#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
```

First we must translate the metadata of our datasets to any of the standard serialisations (file formats) of the World Wide Web Consortium's Resource Description Framework definition, which allows the connection of data across the open internet. At the time of writing this report, the EU Open Data Portal was changing its backend, and for testing purposes, we worked with a dataset from the background of the Open Music Europe project (which had been earlier published by Reprex on Zenodo under the title *The turnover of the ration broadcasting industry in Europe*.)

The dataset itself cannot be downloaded from a data catalogue. It is an abstract intellectual work, similar to musical work or a literary work. A musical work is accessible in printed sheets or recordings, and a dataset in a distributed data file.

```
<https://doi.org/10.5281/zenodo.5652118>
    <a> "dcat:Dataset" ;
    <dcat:distribution> <https://zenodo.org/records/5652118/files/codebook_trb.csv>, "http
    <dct:creator> <https://orcid.org/0000-0001-7513-6760> ;
    <dct:description> "\"The turnover of the ration broadcasting industry in Europe.\"@en"
    <dct:identifier> <https://doi.org/10.5281/zenodo.5652118> ;
    <dct:issued> "2022-06-03T00:00:00Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
    <dct:modified> "2022-06-04T00:00:00Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
    <dct:publisher> <https://isni.org/isni/000000050973936X> ;
    <dct:title> "A rádió szektor forgalma Európában\"@hu", "\"Turnover of the Radio Broadc
    <edp:originalLanguage> <rdf:resource> <http://publications.europa.eu/resource/authorit
```

We can provide further provenance information about the dataset; in production, we will provide information on software agents (tools) used, researchers, data managers and curators and their organisations involved. As a bare minimum, we provide machine-readable information about the technical publisher of the dataset, Reprex B.V:

```
<https://isni.org/isni/000000050973936X>
    <a> "foaf:Agent" .
```

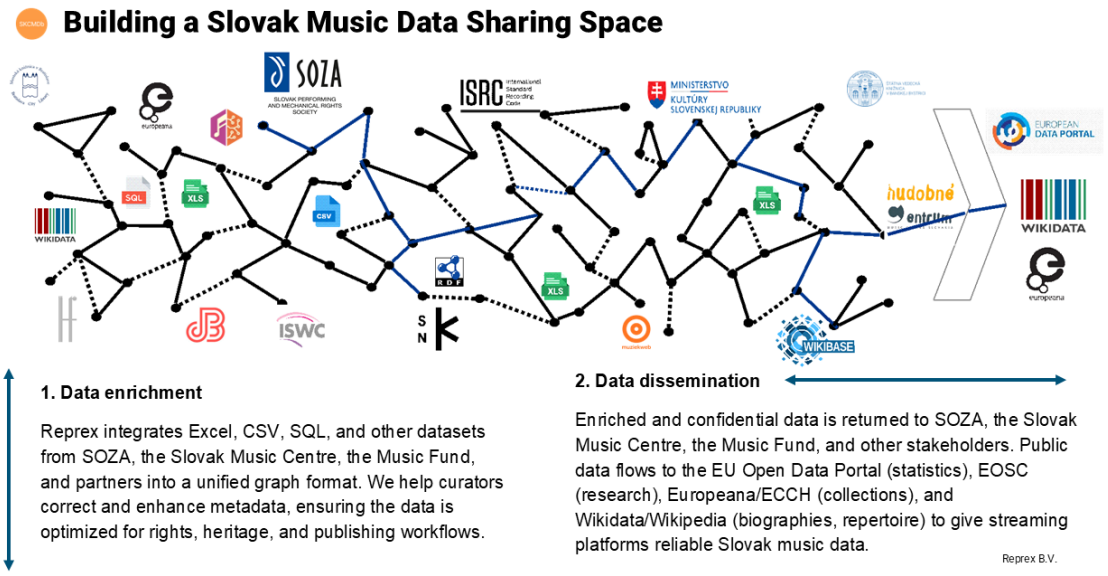
And then we point the user the downloadable files (distributions) of the dataset with the rights statements and licenses. We use the Creative Commons CC BY 4.0 license, similar to Eurostat on the EU Open Data Portal, and we state that the dataset is open for the public.

```
<https://zenodo.org/records/5652118/files/codebook_trb.csv>
    <a> "dcat:Distribution" ;
    <dcat:accessURL> <https://zenodo.org/records/5652118/files/codebook_trb.csv> ;
    <dcat:byteSize> "41672" ;
    <dcat:downloadURL> <https://zenodo.org/records/5652118/files/codebook_trb.csv> ;
    <dcat:mediaType> "text/csv" ;
    <dct:license> <http://publications.europa.eu/resource/authority/licence/CC_BY_4_0> ;
    <dct:rights> <http://publications.europa.eu/resource/authority/access-right/PUBLIC> ;
    <owl:sameAs> <https://zenodo.org/records/5652118> .
```

3.3.2 Documentation

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

3.4 Disseminate



3.4.1 Open Music Observatory

The Open Music Observatory website provides

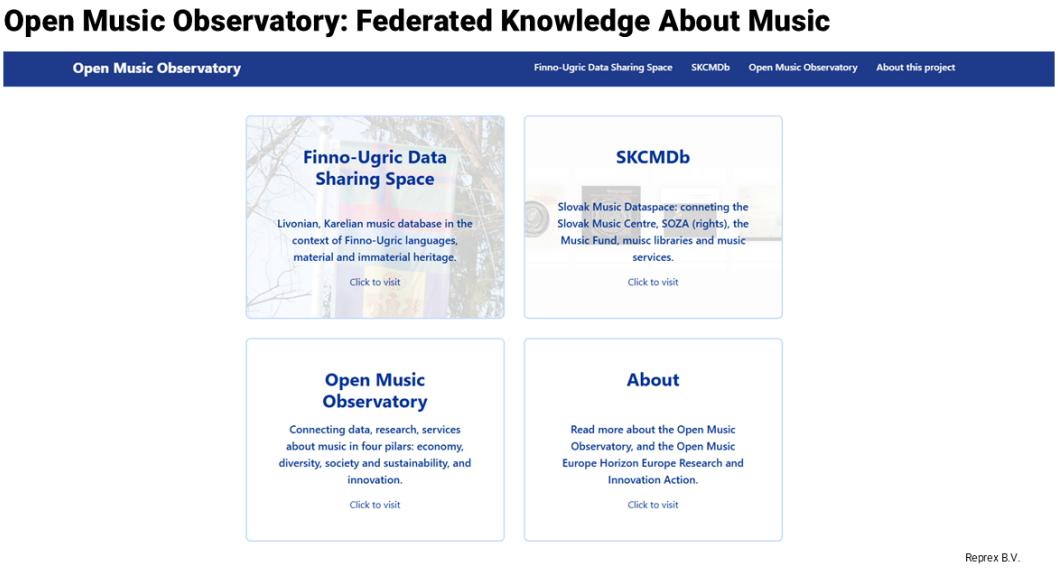


Figure 3.3: Our temporary landing page on <https://dataobservatory-eu.github.io/omo-landing-page>

The Open Music Observatory is a federated dataspace with joint services. The temporary data sharing space currently contains three, and soon four federeated spaces.

- The **Finno-Ugric Data Sharing Space** is our experimental data sharing space that connects the music and broader immaterial and material heritage of fifteen small, mostly endangered ethnolinguist groups; including our Livonan Music Database (see Annex), Karelian Music Database, contemporary and heritage Mari, Udmurt, Székely, Csángó and other music. https://finnougric.net/en/
- The **Slovak Comprehensive Music Database** connects the services of the rights management agency SOZA, the Slovak Music Center, the Slovak National library and some public libraries of Slovakia. It is gradually filled up with music, and it shows where Slovak music (in the form of music recordings, videos, printed sheet, or other publications) can be found and listened to on streaming services, libraries, webshops. This federated unit is well-governed by Slovak national entities. https://hudobnadatabaza.sk/en/
- The **Open Music Observatory** is mainly intended to give access to the Economy, Diversity, Society and Sustainability and Innovation data and publications of the Open Music Europe (OpenMusE) project as well as data from Poland, Latvia, Portugal and other countries. This federated unit is not yet finished by the OpenMusE consortium and is designated to be handed over to pan-European organisations of music http://135.181.91.51:3008/en/

The **Hungarian Comprehensive Music Database** will follow soon.

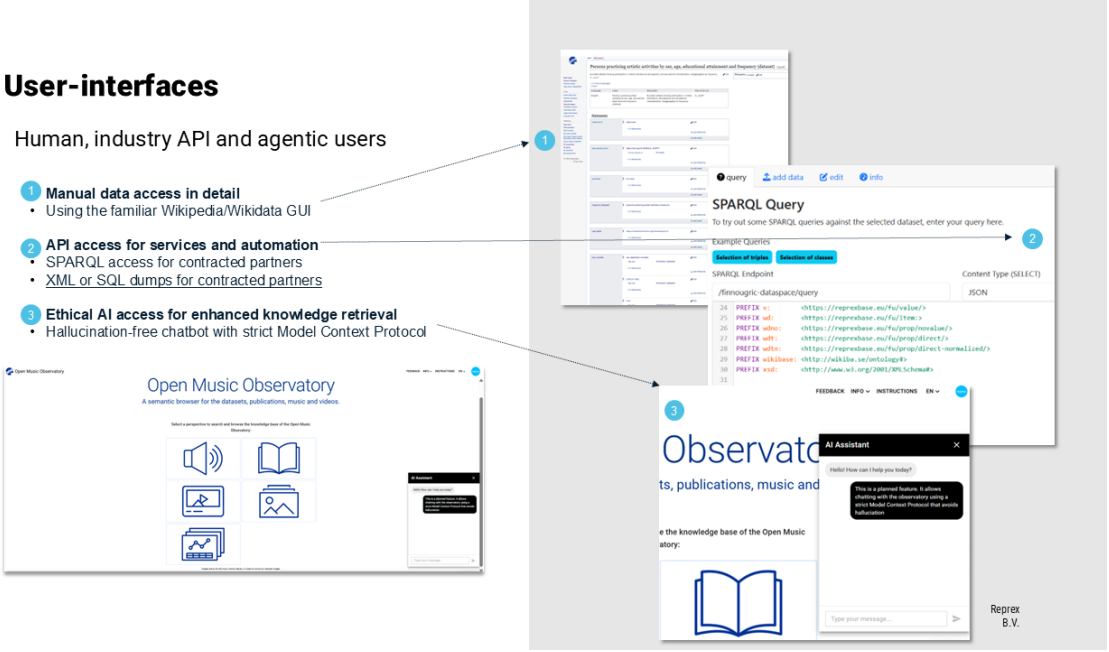
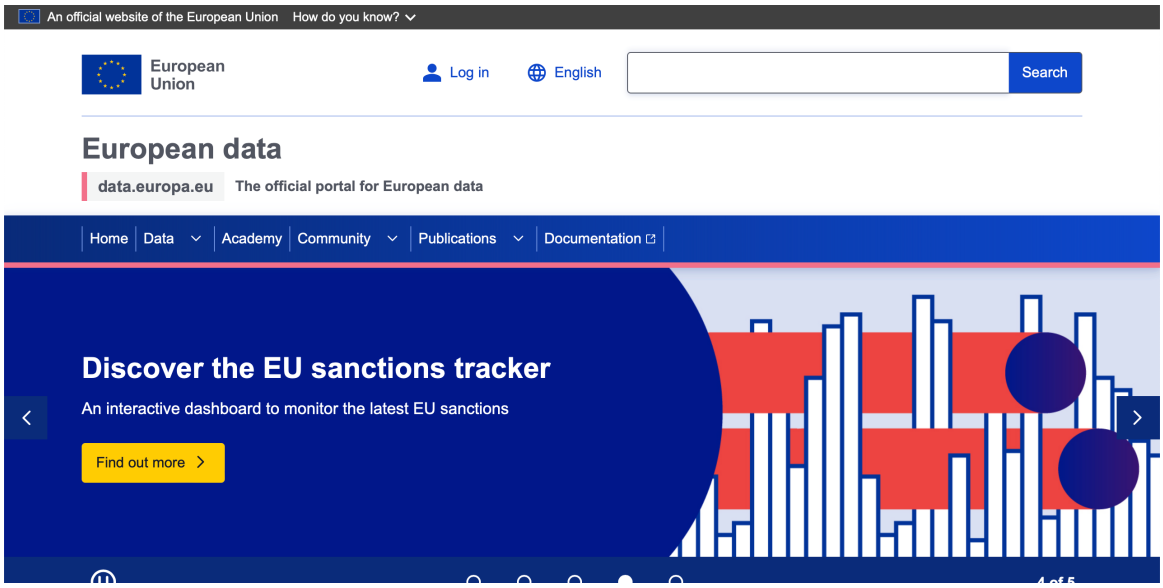


Figure 3.4: Services for individuals, API access for industry and research partners, and planned trustworthy AI.

The Observatory, unlike, for example, the European Audiovisual Observatory, has no permanent staff, and it is completely automated. It offers four layers of services.

- Access to the Wikibase Suite GUI that gives very detailed description of datasets, publication, musical works and their textual or recorded, video manifestations. (See for example: fudss:Q4745) Manual access and ad hoc dumps for RDF serialisations of this information (See for example Q4745.ttl, or Q4745.jsonld.)
- Access to the Sampo open-source semantic browsers that gives faceted search access for researchers, individuals, who cannot write API queries. See for example the https://hudobnadatabaza.sk/en/
- API access to partners (requires password, authentication) that provides SPARQL queries with CSV download.
- A planned extension for December-January is a trustworthy AI chatbot that connects AI agents via a strict Model Context Protocol to the knowledge base(s) of the observatory, and allows custom knowledge retrieval from the knowledge base without hallucations in natural language.

3.4.2 EU Open Data Portal

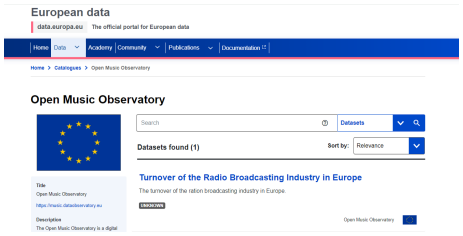


The portal is a central point of access to European open data from international, European Union, national, regional, local and geodata portals. It consolidates the former EU Open Data Portal and the European Data Portal.

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).

3.4.3 Europeana Integration

We promised interoperability with Europeana, and after a lengthy process Reprex concluded a *Data Exchange Agreement* with **Europeana Sound** (the music and sound aggregation center of Europeana hosted by the British Library's Sound and Vision area) and sent pilot data from Latvia.

3.4.4 European Collaborative Cloud for Cultural Heritage

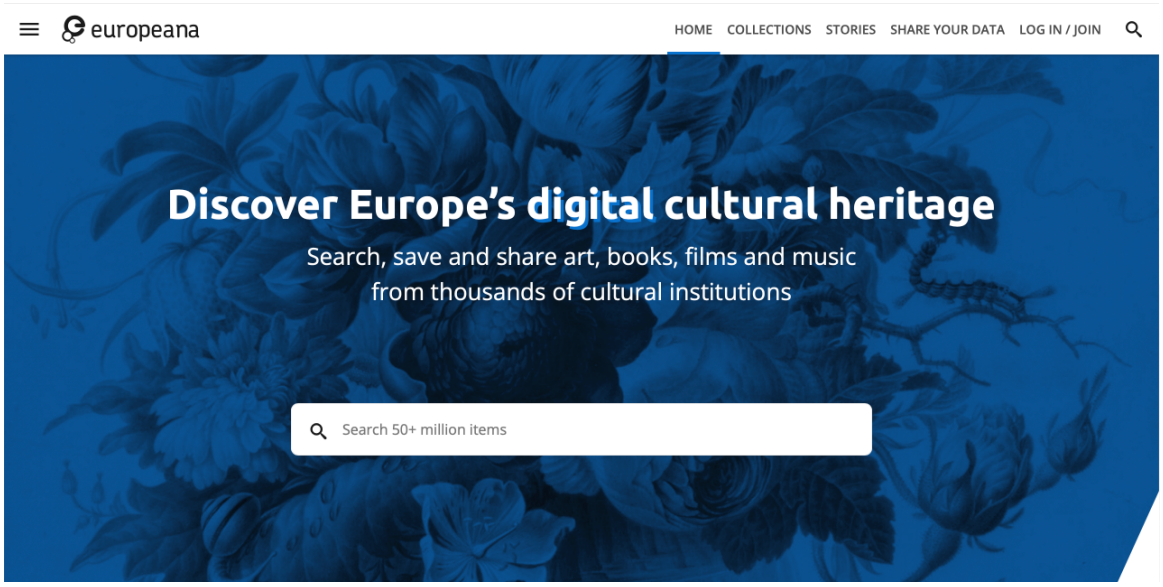


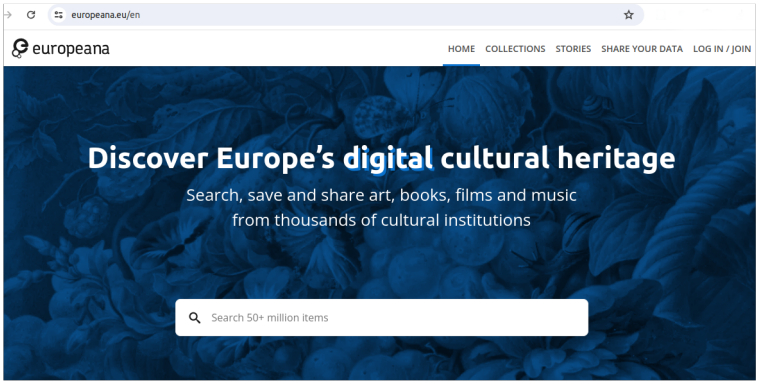
Figure 3.5: Europeana is a bottom-up, decentralised data aggregator for the cultural heritage part of the music sector and the broader cultural and knowledge sector.

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 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 Collaborative Cloud for Cultural Heritage, which has less restrictive data licensing policies, about a more extensive dissemination point for our collection datasets.

> **Note**
>
> More information about our services for curators of music collections and smaller independent repertoires can be found on our website .

3.4.5 European Open Science Cloud

The ambition of the European Open Science Cloud (EOSC) is to provide European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and reuse data, tools and services for research, innovation and educational purposes. Naturally we want to ensure that users of the Open Music Observatory participate in these cloud services, either as research providers or as research users. The EOSC is recognised by the Council of the European Union among the 20 actions of the policy agenda 2022-2024 of the European Research Area (ERA) with the specific objective to deepen open science practices in Europe. It is also recognised as the "science, research and innovation data space" which will be fully articulated with the other sectoral data spaces. Given that the OMO is also following a data space architecture that is

desigend to follow the European Interoperability Framework, the Open Music Observatory can be federated with, and can fully work with the EOSC.

The Open Music Observatory is connecting to the EOSC via two key services of `OpenAIRE`. OpenAIRE itself is a Non-Profit Partnership of 50 organisations, established in 2018 as a legal entity, OpenAIRE A.M.K.E, to ensure a permanent open scholarly communication infrastructure to support European research, and it is a key implementer of the European Open Science Cloud. Or connection to OpenAIRE services guarantee our full compliance and use of EOSC.

A key partner of OpenAIRE is CERN, which is manages the Zenodo open library and repository. We rely on the services of Zenodo for document identification, long-term archiving, and offering immediate access to our statistical datasets, visualisations, reports and other library-ready products. Similarly to the forming `EU Open Research Repository`, a Zenodo-community dedicated to fostering open science and enhancing the visibility and accessibility of research outputs funded by the European Union, managed by CERN on behalf of the European Commission, we have created a similar `Open Music Observatory Repository` on the platform. Our repository is fully interoperable with the `EU Open Research Repository` (in pilot phase in June 2024) and with the entire EOSC.

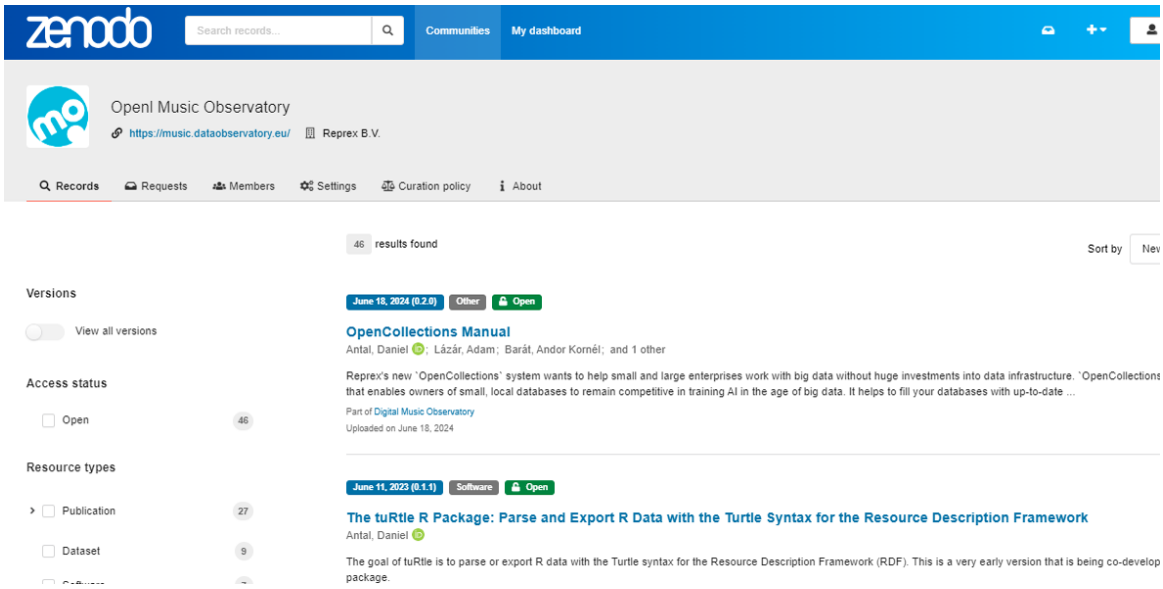


Figure 3.6: For interoperability with EOSC and OpenAIRE and for long-term storage we store each distribution of the datasets among publications, documents on Zenodo, too, in the Open Music Observatory community.

The semantic service of OpenAIRE, the `OpenAIRE Graph` is a collection of interlinked research objects that aggregates metadata records from more than 70K scholarly communication sources from all over the world for researchers, service providers, research managers and policy makers, by following a participatory approach.

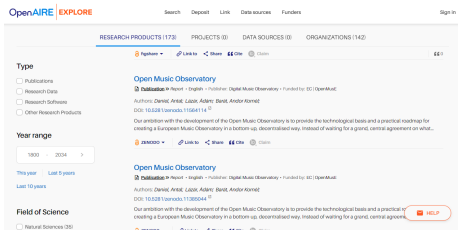


Figure 3.7: For interoperability with EOSC and OpenAIRE and for long-term storage we store each distribution of the datasets among publications, documents on Zenodo, too, in the Open Music Observatory community.

3.5 Metadata

Europeana publishes the metadata in Turtle serialisation. The Open Music Observatory will provide access to the metadata in TTL and CSV distributions.

3.5.1 Wikibase & Wikidata

Our main dissemination point for microdata are two Wikibase Suite installations, one for the Slovak Comprehensive Music Database, and one for other music.

- For data related to Slovakia in the `Slovak Comprehensive Music Database`, we provide access at <https://reprexbase.eu/skcmdb/>.
- For all other data we provide access at https://reprexbase.eu/openmusic/; this data has a less complex data management and governance. This access point provides access for individual datasets.

We may use the openmusic.wikibase.cloud. Wikibase Cloud is an initiative of Wikipedia and Wikidata to bring more specialised data into the Wikimedia ecosystem. It may be a staging area for harmonisation with Wikidata.

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.

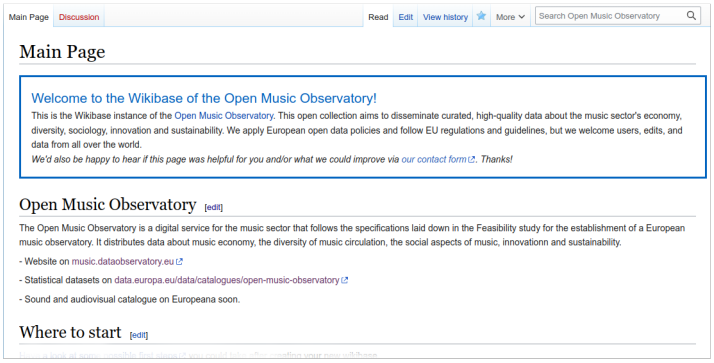
Wikidata itself is a gigantic *Wikibase instance*. Their user interface is similar, but depending on what the administrator of your Wikibase instance allows you to do, you are likely to have more freedom to edit certain elements, like properties, than on Wikidata. Wikidata must protect the integrity 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.



3.5.2 Music Observatory Website

> ⚠ Warning
>
> We will completely revamp the website before submission and provide here a short overview with screenshots.

3.5.3 API Endpoint

> ⚠ Warning
>
> We will provide here a linked screenshot of our API endpoint before submission

4 Architecture

The *Open Music Observatory* (OMO) is designed as a *decentralised, federated, semantically interoperable knowledge infrastructure*. Its architecture is the direct consequence of the policy, governance, and methodological requirements described in the Background chapter (see Chapter 2, Section 2.3, Section 2.4, and Section 2.5). The architecture is not an aesthetic or purely technical choice: it is the *only sustainable and legally compliant way* to realise the "open, scalable data-to-policy pipeline" mandated in the Grant Agreement. The architectural choice is also reinforced by the findings of the *CITF First Project Report*, which examined the requirements for trustworthy, lifecycle-aware copyright infrastructures in the AI era. CITF identifies the same structural needs—federated governance, interoperable identifiers, machine-readable rights metadata, and verifiable provenance—that underpin OMO's design. This convergence shows that the Observatory's dataspace architecture is not only technically justified but part of a broader European shift toward distributed, standards-based copyright and metadata infrastructures.

Traditional observatories created in the 1990s and 2000s were built on *centralised databases*, static data submissions, and annual reporting cycles. Such architectures are no longer viable in an ecosystem where:

- metadata is produced continuously across incompatible systems;
- rights, repertoire, event, and business data change daily;
- cultural-heritage and community archives hold essential non-market metadata;
- statistical offices and ministries have different identifiers and legal mandates;
- AI-assisted workflows require transparent provenance and versioning;
- multilingual and cross-domain reconciliation is unavoidable.

These conditions make a *federated dataspace* (see Section 2.4) the only feasible approach. Within such a dataspace, **Wikibase/Wikidata** provides the most suitable technological foundation.

4.1 Knowledge Base for Humans and Machines

Knowledge representation describes how information about a given domain is structured so that humans and machines can interpret it in a consistent way. In the language of ISO standards, a knowledge representation identifies the concepts that matter, how these concepts relate to one another, and how they are recorded in a data model. In complex domains such as music production, rights management, or cultural statistics, knowledge representation acts as the bridge between raw data and meaningful analysis.

A `knowledge base` is the concrete form of this representation. It is a structured collection of facts, identifiers, rules, and relationships that model a specific domain and support reasoning over it. According to ISO definitions, a knowledge base may hold human-created mappings, machine-generated inferences, and rules that encode expertise from the various aspects of live music, recorded music, music education, cultural and music policies, music tech innovation, sustainability reporting and other areas.

A *shared knowledge base* is necessary because the data that describe film production are distributed across many actors and formats.

- Live and recorded music producers maintain accounting ledgers and project-based cost codes.
- Public authorities collect administrative data for tax rebates, audits, and cultural policy.
- Statistical offices use NACE and CPA classifications, supply-use tables, and environmental accounts to represent the wider economy.

These systems describe overlapping parts of the world but in their own conceptual schemes and with their own definitions.

4.1.1 Shared Services: European Interoperability Framework

The *European Interoperability Framework* recognises that in such situations, achieving interoperability requires a shared space where meaning can be aligned without centralising all data. This is why we design MMAT as a *data sharing space* rather than a single repository. The data remain with the actors who produce or control them, but the shared knowledge base provides common identifiers, mappings, and semantic patterns that allow them to communicate.

A *data sharing space* solves a structural problem, which is explained further in the Chapter 5[1].

- Live and recorded music productions use project-specific terminology and internal codes.
- Accounting systems are arranged around financial categories.
- National accounts rely on NACE activities and CPA products.
- Environmental accounts follow the structure of air emissions accounts and input–output extensions.

None of these systems alone is sufficient to answer cross-cutting questions about economic impact, cultural policy, or sustainability. Without a shared data space, each actor must reconcile data manually, leading to inconsistencies, duplicated effort, and results that cannot be compared across projects or years.

[1]See in particular (Kung, Walshe, and Wenning 2023) and the Chapter 5 for more details.

A well-designed knowledge base captures the relationships among these systems and makes them reusable. It also makes it possible to combine administrative records with statistical classifications without violating legal boundaries around data protection or commercial confidentiality.

Interoperability is essential because screen production depends on many autonomous systems that must exchange information in a trustworthy and legally compliant manner. - Granting authorities need accurate and consistent metadata to evaluate public-funded projects and report aggregated results. - Producers need a simple way to reuse their accounting data for financial control, audit, and sustainability reporting. - Emerging sustainability requirements in Europe and the United States increasingly require value-chain information, including emissions embedded in purchased goods and services.

These requirements cannot be met if each system remains isolated. Semantic interoperability allows systems to describe their data in a shared language, even if each retains its own internal structures.

Technical interoperability ensures that these mappings can be exchanged across platforms.

Legal interoperability clarifies what can be shared and how data protection obligations are met. The European Interoperability Framework stresses that all three layers are necessary to enable cross-administration and cross-sector data flows.

The architecture of Open Music Observatory's system, ReprexBase, adapts these principles for the music sector production. The music data sharing spaces of the OMO are not single databases. They form a a federated structure built around a common knowledge base.

The knowledge base contains the standard identifiers, mappings, and controlled vocabularies that describe the film domain in a machine-interpretable form.

These include mappings between chart-of-accounts categories, mandatory film-production cost codes, and national statistical classifications such as NACE and CPA. They also include alignments to supply-use and input–output tables, cultural satellite accounts, and environmental accounts. This allows a production ledger to be connected to national economic statistics through well-defined transformations. It also allows emissions accounting to be computed using harmonised environmental intensity data.

The architecture combines several layers.

At the **foundational layer**, Open Music Observatory's central, supranational module maintains reference identifiers, domain vocabularies, and canonical mappings. This is where the authoritative concepts are defined, drawing on ISO terminology, national classifications, and legal definitions.

The **semantic layer** contains the patterns and equivalences that link different institutional models. It allows film cost codes to be interpreted as economic activities, or accounting categories to be reconciled with national accounts. Because different institutions organise their data according to different logics, this semantic layer is deliberately modular, following the approach we have used in other cultural sectors. It connects, rather than replaces, existing models.

At the **technical layer**, MMAT uses interoperable interfaces and automated pipelines to exchange data across systems. This is necessary for integrating ERP exports, digital audit files, statistical registers, and environmental accounting tools.

4.1.2 Fixing Metadata At Source; But Also Work with Legacy Metadata

This architecture supports both preventive and curative data workflows.

1. **Preventive workflows** make data interoperable at the moment of creation by assigning identifiers, validating codes, and capturing provenance–this is what the European Parliament's resolution calls for referring to the future: that future music assets should we well described.
2. **Curative workflows** repair legacy data, reconcile inconsistent categories, and align historical records with current classifications.

In the music domain, both are necessary, because the copyright and neighbouring right protection terms necessitate the rights management and documentations up to about ~130 years (70 years after the passing of the last surviving composer.) This means that our systems must carry the data of the entire production catalogue of the 20th century and the first quarter of the 21st century.

Studio productions as well as live productions, particularly music festivals generate large volumes of financial and descriptive data in short time windows, and much of this data was not originally designed for reuse in national or environmental statistics. A knowledge base allows these datasets to be repaired and mapped in a consistent and transparent way.

The need for such architecture is already visible in the music sector, where fragmented metadata and incompatible workflows create recurring reconciliation costs. Our *Green Paper on AI, Data Governance, and Metadata Policies for Europe's Music Ecosystem* demonstrates how a federated data sharing space can connect public and private actors and enable cross-sector analysis while respecting legal and institutional boundaries[2]. It also shows that centralisation is unrealistic in a domain where autonomy, specialised workflows, and legal mandates differ across actors.

Reprexbase, and the core Open Music Europe module of our Observatory therefore provides a stable foundation for current and future needs.

- It supports economic impact analysis by connecting production data to national accounts. It supports cultural policy by linking productions to cultural statistics and sectoral indicators.
- Its diversity module helps to monitor the local content regulations of many countries (that require certain local, national or European production to be present in the broadcast stream) or KeyChange gender equity pledges.

[2]This green paper is the accompanying policy document of this technical description (Antal 2025f).

- It supports environmental analysis by integrating emissions accounting into the same data pipelines.It prepares the sector for future sustainability reporting obligations by establishing harmonised and reusable mappings.

In this sense, the knowledge base does not serve a single reporting purpose. It creates a shared language that enables producers, public bodies, and researchers to draw consistent and verifiable conclusions from heterogeneous data sources.

By grounding the architecture in open standards and the principles of the European Interoperability Framework, the *Open Music Observatory's* `ReprexBase` system becomes a long-term infrastructure rather than a single project database. It creates a space where data can be shared meaningfully, legally, and reproducibly, and where new analytical tools can be integrated without restructuring everything from scratch. This ensures that the film sector can adapt to new policy requirements while retaining control over its own data and workflows.

4.2 Why Wikibase Is the Right Foundation for the Open Music Observatory

The European music ecosystem produces data in many incompatible formats:

- repertoire and rights databases (CMOs, publishers, labels),
- cultural-heritage and performing-arts collections,
- company registries and economic statistics,
- event metadata from festivals and venues,
- community-maintained sources (Wikidata, folk archives, local heritage groups).

These sources use different identifiers, legal definitions, languages, and metadata schemas. A *knowledge graph* is therefore indispensable: no relational or document database can reconcile these sources while maintaining provenance, multilinguality, and entity-level linking.

> **i** Alignment with CITF's Copyright Infrastructure Model
>
> The **CITF First Project Report** defines three layers that a modern copyright infrastructure must satisfy: a *foundational identifier layer* (authoritative PIDs and registries), a *semantic layer* (shared meaning and mapping across domains), and a *technical layer* (APIs, services, resolution, and provenance). Wikibase operationalises this same structure in practice. Its support for persistent identifiers, ontology alignment, multilingual semantics, and transparent versioning makes it fully compatible with the CITF model and positions the Observatory as a concrete, domain-specific implementation of that broader European framework (Partanen et al. 2025).

Wikibase is adopted because it has already proven its suitability in domains facing the same structural issues as the music ecosystem: fragmented identifiers, inconsistent authority control, multilingual metadata, cross-domain vocabularies, and parallel institutional workflows.

4.2.1 Proven in real-world scenarios highly similar to music

Wikibase is widely used by libraries, archives, museums, national cultural bodies, and open-science infrastructures. These institutions face the same challenges OMO addresses:

- reconciliation of people, works, events, organisations, and places;
- multilingual labels and aliases;
- authority file alignment (ISNI, VIAF, ORCID, GND, BNF, corporate registers);
- provenance tracking and version history;
- SPARQL-based validation and constraint checking.

The *GLAM-Wiki ecosystem*, national knowledge graphs, and EU-funded linked-data projects have collectively demonstrated that Wikibase is an effective intermediary between:

- authoritative PID systems;
- domain ontologies (CIDOC-CRM, RiC-O, DDI, DCAT);
- community-curated knowledge models.

This track record gives OMO a mature, future-proof, and standards-aligned foundation.

4.2.2 Already aligned with Europe's digital knowledge infrastructure

Wikibase aligns with existing institutional practice across Europe. Many major knowledge centres and initiatives already use Wikibase/Wikidata:

- national libraries and archives,
- national cultural-heritage aggregators,
- research infrastructures,
- public-sector linked-data programmes,
- the *EU Knowledge Graph* initiative.

Adopting Wikibase ensures that the *Open Music Observatory* fits directly into the *European interoperability ecosystem* (see Section 2.3).

> Footnote: See *Wikibase as an Infrastructure for Knowledge Graphs: the EU Knowledge Graph* (Diefenbach, De Wilde, and Alipio 2021) and (2020 2020).

4.2.3 Demonstrated support for required OMO functionality

Everything OMO needs has already been demonstrated in production Wikibase environments:

- *authority control* for creators, ensembles, organisations, venues;
- *multilingual and multiscript modelling* for names, places, works;
- *cross-domain entity linking* (work–recording–performance–rights–heritage);
- *event-based and entity-based models*;

- *SPARQL validation*, schema constraints, and automated reconciliation.

This means OMO does *not* invent an untested paradigm: the consortium integrates proven practices from:

- national registries,
- performing-arts knowledge graphs,
- the Slovak pilot and Finno-Ugric metadata federations developed inside the project.

4.2.4 Fits EU policy preference for open-source and trustworthy AI

Wikibase is open-source, auditable, and non-proprietary.
It aligns with:

- the EU's preference for *open-source digital public infrastructure*,
- FAIR and CARE principles,
- trustworthy AI requirements (provenance, transparency, versioning),
- cross-border interoperability mandates,
- decentralised data governance models.

This makes it compatible with Europeana, EOSC/ECCCH, DCAT-AP, and the *European Interoperability Framework*. It is used in EU organisations, too. [3]

4.2.5 The most widely used graph-editing interface in the world

Tens of thousands of data stewards, librarians, researchers, and citizen-scientists already know how to edit Wikibase/Wikidata.

This provides OMO with:

- an immediate user base,
- a ready-made contributor community,
- institutional familiarity across Europe,
- workflows already adopted in GLAM and research sectors.

No alternative open-source system has remotely this level of adoption.

[3] On official adoption: EU Knowledge Graph (Diefenbach, De Wilde, and Alipio 2021); SEMIC guidelines (SEMIC Support Centre 2023).

4.2.6 A hybrid model that fits real institutional workflows

Wikibase uniquely accommodates:

- spreadsheet-based workflows (Excel, CSV),
- relational database exports,
- statistical microdata reference linking,
- complex semantic modelling,
- API-based ingestion,
- R and Python pipelines.

It is a *practical compromise* between triple stores, document databases, and relational systems — perfect for a music ecosystem where many partners still rely on basic tools. It has proven to be useful in music services[4], and more generally on small- and large scale European knowledge institutions (national libraries, libraries, archives, museums[5].)

4.3 How Wikibase Fits into the Open Music Europe Data-to-Policy Pipeline

The data-to-policy pipeline defined in the Grant Agreement and documented in the Background chapter (see Section 2.5) provides the methodological backbone of OME. Wikibase is the component that makes this pipeline operational.

4.3.1 Wikibase supports each stage of the pipeline

4.3.1.1 Data collection

- imports from Excel, CSV, SQL, APIs, and legacy systems;
- immediate linkage to persistent identifiers;
- entity reconciliation as part of ingestion.

4.3.1.2 Validation and reconciliation

- authority-control workflows for people, works, organisations, and places;
- constraint-based quality checks;
- SPARQL-driven validation;
- alignment with external authority files.

[4] On Belgian pilots: *MetaBelgica* (Stallmann et al. 2023) and Flemish performing arts enrichment (Magnus and Van D'huynslager 2021).

[5] See for example: On Wikidata/Wikibase in heritage: (Bianchini, Bargioni, and Pellizzari di San Girolamo 2021; Sardo and Bianchini 2022).

4.3.1.3 Harmonisation and enrichment

- multilingual labels and roles;
- event-based and relationship-based modelling;
- addition of contextual metadata by different institutions;
- integration of domain vocabularies.

4.3.1.4 Activation for analysis

- SPARQL endpoints for programmatic access;
- JSON-LD, RDF dumps, and REST APIs;
- R and Python pipelines use stable URIs for reproducibility.

4.3.1.5 Indicator construction

- cross-domain indicators linking economic, cultural, rights, and heritage data;
- entity-level referencing ensures indicators are traceable and verifiable.

4.3.1.6 Interpretation and contextualisation

- experts review, annotate, and correct metadata through a human-readable interface;
- provenance guarantees transparency.

4.3.1.7 Policy translation and observatory outputs

- live, federated knowledge base powering the OMO front end;
- entity profiles, metadata dashboards, and linked methodological documentation;
- direct links from indicators to source entities and datasets.

4.3.1.8 Feedback loop

- corrections flow back into the shared graph;
- updated authority files update all downstream indicators;
- the observatory improves over time.

4.4 Summary

Wikibase is adopted because it best fulfils the policy requirements, semantic needs, technical constraints, and governance expectations described in the Background chapter:

- It is *proven* in domains identical to ours.
- It is *aligned* with the EU's open-source, dataspace, and interoperability agenda.
- It is *widely adopted* by the very institutions we must interoperate with.
- It *supports the entire Open Music Europe data-to-policy pipeline.*
- It allows the Observatory to operate as a *decentralised, federated, evolving knowledge infrastructure.*

This architecture is therefore not an optional design preference but the *only viable model* for delivering the European Music Observatory described in the Grant Agreement.

5 Data coordination

Cultural and music data in Europe is generated and maintained by a highly diverse set of organisations. National libraries, archives, museums, collective rights management organisations, academic research projects, streaming platforms, local cultural associations, and commercial distributors all describe the same creators, the same works, and the same recordings—but they do so for different purposes, using different conceptual models, and under different legal and organisational constraints. These systems have grown organically over decades, reflecting local traditions, professional cultures, and technical possibilities. As a result, the same piece of music may be catalogued differently in an archive, credited differently in a rights database, indexed differently in a library catalogue, and published differently on a digital music service.

This fragmentation was already identified in the *Feasibility Study for a European Music Observatory*, which noted that music-sector metadata is "structurally scattered across private and public registers that seldom interconnect" (European Commission et al. 2020, 9–10). The Study explicitly highlighted the absence of a "functional link between ISWC, ISRC, ISNI and VIAF" as a core obstacle to evidence-building [p. 46–48]. The *CITF First Project Report* reaches the same conclusion in the textual domain, observing that rights metadata lacks "cross-registry alignment, lifecycle-based provenance, and interoperable identifiers" [Partanen et al. (2025), p. 20–22; p. 31]. These findings reinforce the need for a coordination model that can operate across independent registers without replacing them—precisely the purpose of a data sharing space.

This fragmentation is not a sign of failure; it is the natural consequence of cultural stewardship being distributed across many institutions with distinct missions. However, fragmentation becomes a problem when people attempt to connect, reuse, or reinterpret information across domains. Without coordination, data cannot be trusted across systems: names do not match; contributor roles are incompatible; identifiers are missing or ambiguous; and important aspects of minority or community-based cultural heritage may be lost or become invisible.

The need for coordination arises from this structural diversity. Institutions should not be forced to adopt a single metadata standard or a single software system, but they must be able to communicate meaningfully. Coordination provides the minimum shared semantic and technical foundations that enable such communication. Coordination also supports equity and cultural diversity. Smaller institutions—community archives, ethnographic collections, minority cultural groups, and local music organisations—often lack the resources to operate modern digital infrastructures. Without coordinated frameworks, their data cannot easily participate in larger European infrastructures or data spaces, which in turn reinforces cultural imbalances.

Coordination therefore serves not only technical efficiency but also cultural policy objectives related to inclusion, multilingualism, regional diversity, and the representation of endangered traditions.

The European Parliament's *Resolution on Cultural Diversity and the Conditions for Authors in the European Music Streaming Market* makes this a policy priority, stressing that fair remuneration and discoverability depend on "comprehensive and accurate allocation of metadata from the time of creation" and on adoption of all international identifiers (IPI, ISWC, ISRC, ISNI, IPN) (European Parliament 2024, recital 32). The EU *Music Ecosystem Study* similarly notes that fragmentation in metadata and identifier systems "prevents cross-border comparability and weakens the evidence base for policy" (Music Moves Europe 2024, 42–45). Coordination is therefore a legal and policy requirement, not simply a technical improvement.

5.0.1 The European Interoperability Framework (EIF)

The *European Interoperability Framework* (EIF) provides guidance for how heterogeneous public-sector systems can work together without sacrificing their autonomy. It identifies four layers of interoperability—legal, organisational, semantic, and technical—and emphasises that meaningful interoperability requires all four to be addressed[1].

The EIF provides the most relevant cross-sectoral guidance for this problem. The *European Interoperability Framework Implementation Strategy* stresses that interoperability must balance institutional autonomy with cross-sector alignment, and that semantic mediation—not schema unification—is the correct method for heterogeneous domains (European Commission 2017b, 9–14). This perspective is echoed in the *Data Spaces Support Centre Blueprint v2.0*, which states that common European data spaces must provide "shared semantics, identifier resolution, and rulebooks for interoperability" rather than imposing a single ontology (Data Spaces Support Centre 2025, 35–39).

The EIF's underlying philosophy is that institutions will always differ in their practices, constraints, and goals. Therefore, interoperability cannot be achieved by imposing uniformity. Instead, it must enable institutions to maintain their internal systems while still participating in a shared ecosystem.

Although the EIF was not developed specifically for the cultural or music domains, but generally to public digital services, its principles apply directly, because most music services are digital, and they are often public in nature, too.

[1]The EIF defines layered interoperability (legal, organisational, semantic, technical) (European Commission 2017a). The *European Strategy for Data* frames subsidiarity as compatible with federation (European Commission 2020). BDVA and the Federation Working Group emphasise that interoperability frameworks are needed to operationalise federation (BDVA/DAIRO 2023; BDVA/DAIRO Federation Working Group 2023).

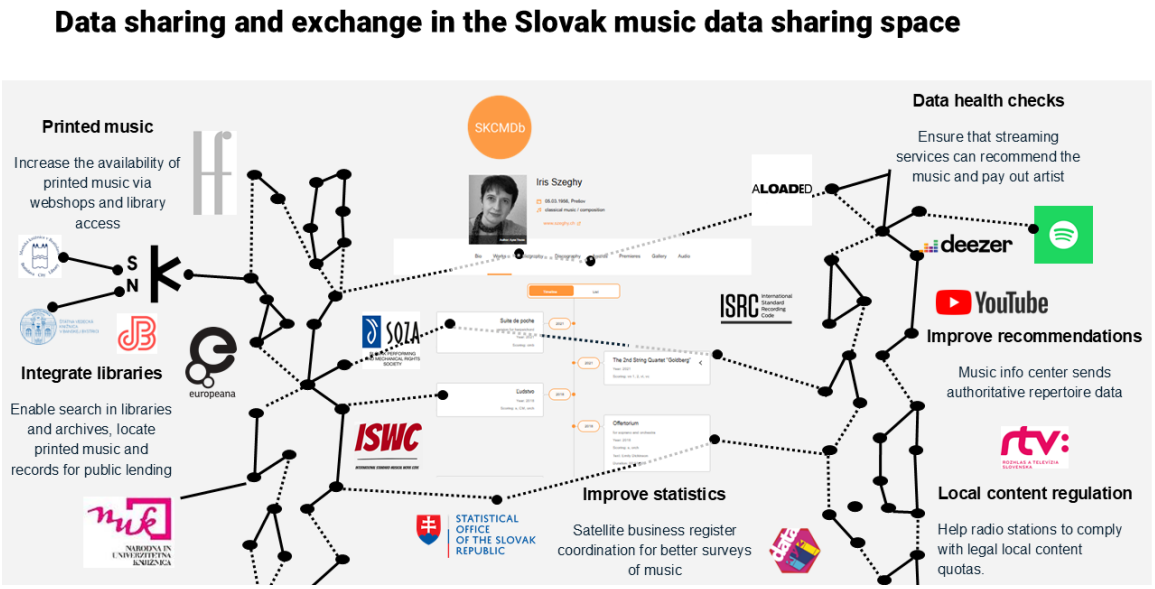


Figure 5.1: The role of data interoperability is not to exchange or centralise data, but to create digital services in libraries, rights management, distribution, streaming platforms that work together well. For the service providers they should work with less data processing cost, and for the user they should give a higher value or better user experience.

Cultural data infrastructures face the same challenges the EIF describes:

- different legislation (e.g., copyright, orphan works, deposit laws),
- different organisational workflows (cataloguing, digitisation, rights clearance),
- and different semantic traditions (archival provenance, library authority control, rights metadata, music industry credits).

The EIF also emphasises reusability, transparency, and proportionality—all central issues for cultural data governance, particularly when dealing with sensitive material, intangible heritage, and community-generated knowledge. Most importantly, the EIF recognises that interoperability is not a technical property alone.

It is a sociotechnical negotiation across institutions, requiring trust, shared agreements, and mechanisms for maintaining alignment over time. This recognition makes the EIF a strong conceptual foundation for coordinating cultural and music data, even when many actors fall outside the classical public-sector domain.

5.0.2 Extending the EIF to public and private service coordination

Cultural data ecosystems are unique because they span both public and private infrastructures.

A library's bibliographic record for a musical work interacts with rights information managed by collective management organisations, which in turn interacts with ISWC and ISRC registries maintained by industry actors.

A museum's digitised ethnographic collection may be reused by researchers, by minority communities seeking cultural revival, or by streaming services presenting curated heritage playlists. Public institutions depend on commercial metadata pipelines for up-to-date identifiers, contributor roles, and distribution metadata; commercial actors depend on public institutions for authoritative information, contextual enrichment, and long-term preservation.

To coordinate across this mixed ecosystem, the EIF must be extended beyond its original scope.

This extension is explicitly anticipated in the EU's data governance framework. The *Data Governance Act* requires that cross-sector reuse of protected data be mediated through shared governance mechanisms and semantic alignment [Regulation (EU) 2022/868, Art. 8–12]. The *Data Act* reinforces that business-to-government and business-to-business reuse must rely on interoperable identifiers and machine-readable formats [Regulation (EU) 2023/2854, Art. 3–4]. The CITF report further argues that trustworthy AI requires alignment across private registries (publishers, collecting societies) and public infrastructures (national libraries, VIAF, ISNI), because "lifecycle-level provenance cannot be ensured in siloed systems" (Partanen et al. 2025, 101–2).

Public–private coordination requires shared semantics, shared identifiers, and shared rules for linking, even when underlying workflows remain distinct. For example, DDEX roles must be interpretable by library and archival systems, just as archival provenance must be interpretable by music distributors when republishing historical recordings. Similarly, community-generated metadata (such as language revitalisation initiatives or fieldwork collections) must be able to participate in rights and distribution pipelines without being forced into commercial templates that erase cultural meaning.

Extending the EIF in this way does not imply that private systems become part of public digital government infrastructures. It means that coordination mechanisms—identifier mapping, semantic mediation, minimal ontological alignment, provenance tracking, rights documentation—must operate across both sectors. This allows public and private services to interoperate responsibly while preserving their organisational autonomy and legal boundaries. The result is a more inclusive, resilient ecosystem, where cultural data can move across domains without losing meaning or trustworthiness.

5.0.3 Data sharing space

A data sharing space is a distributed, federated infrastructure in which institutions agree on a minimal set of rules for linking and interpreting each other's data.

The concept of a data sharing space is directly aligned with the *Design Principles for Data Spaces* published by the International Data Spaces Association and adopted by the European Commission. Data spaces are defined as "federated data ecosystems within a specific domain, operating on shared semantics, trust frameworks, and rulebooks" rather than as central repositories (Nagel and Lycklama 2021, 7–9). The CEEMID initiative anticipated many of these principles a decade earlier by demonstrating that decentralised linking—rather than centralisation—is the only viable model for cultural and creative sector data (Antal 2020a, 18–22).

Unlike a centralised database, a data sharing space does not require institutions to upload all their data into a single platform or adopt a single schema. Instead, it provides a semantic mediation layer—shared identifiers, alignment patterns, equivalence mappings, and reference ontologies—that allows heterogeneous systems to interoperate.

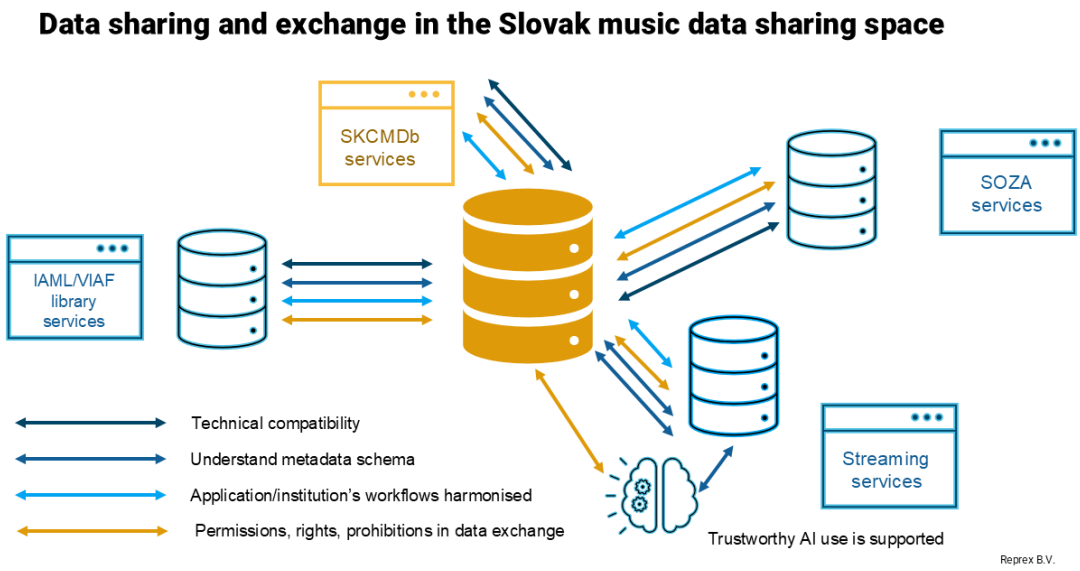


Figure 5.2: A data sharing space is an ideal coordination mechanism for EIF: it ensures that data-driven digital services of libraries, streaming platforms, distributors or collective management organisations can exchange data and data definitions efficiently. See the Annex for more information on the Slovak Music Dataspace behind https://hudobnadatabaza.sk/en/.

Data sharing spaces are particularly suitable for cultural and music ecosystems, where no single actor can control the entire landscape. They allow archives to link to music distrib-

utors, libraries to link to rights registries, and community collections to link to research datasets, without forcing any of these institutions to abandon their internal models.

A data sharing space can support service integration in practical, operational ways. For example, archival metadata describing a traditional song can be transformed into DDEX metadata for legal distribution; a library's authority record can enrich a rights organisation's contributor database; a researcher's MIR dataset can link to digitised collections; and a community can add linguistically and culturally meaningful information without overwriting institutional records.

Because data sharing spaces respect institutional autonomy, they can grow incrementally. Institutions can join at different levels of readiness, linking only what they are comfortable sharing. Over time, shared identifiers and alignment patterns accumulate, improving the quality and connectivity of the ecosystem as a whole.

The data sharing space becomes the place where coordinated services emerge: discovery interfaces spanning multiple institutions, rights-aware access systems, cross-domain research environments, and community-led enrichment workflows. In this way, a data sharing space serves as the structural foundation for a coordinated, equitable, and future-proof European cultural and music data infrastructure.

5.1 Ontologies and Vocabularies in the Open Music Observatory

The **Open Music Observatory (OMO)** does not aim to create new ontologies.
Instead, it integrates and reuses **existing, well-established data models** from the open data and cultural heritage ecosystems.

In practice, this means that our data spaces and pilot projects (e.g. SKCMDb, Finno-Ugric Data Space, TextileBase) combine metadata drawn from:

- **Open Government and Open Science:**
  DCAT-AP — the European data portal metadata model.
- **Cultural Heritage and GLAM:**
  EDM (Europeana Data Model),
  CIDOC CRM, and Records in Contexts (RiC).
- **Emerging Cultural Infrastructure:**
  High-Definition & Collaborative Cultural Data Infrastructure (HDTO).
- **Common Metadata Principles:**
  Dublin Core Terms (DCTERMS) — for general metadata interoperability.

- **Music Industry Standards:**
  ISWC for works,
  ISRC for recordings,
  DDEX for metadata exchange and release categorisation.

Our goal is not to add another layer of complexity, but to **connect these existing models** so that cultural heritage data, rights management systems, and research metadata can interoperate seamlessly.

This bridging strategy mirrors the approach taken in cultural heritage infrastructures such as Europeana (EDM), the emerging Heritage Digital Twin Ontology (HDTO), and the EOSC semantic interoperability guidelines, all of which rely on equivalence mappings and lightweight ontology design patterns rather than top-down schema replacement (ECHOES Ontology Task Force 2025, 5–7; 2020 2020, 13–19). This is the same approach recommended by the CITF report for cross-registry rights metadata (Partanen et al. 2025, 31).

Open Music Observatory Ontology Approach

- The Observatory uses a **Wikibase-based ontology layer**, derived from the Wikibase Ontology.
  This defines classes and properties that are compatible with the Wikidata ecosystem.
- **Equivalence links** (`owl:equivalentClass`, `owl:equivalentProperty`) are added to map our entities to reference ontologies such as:
  - `dcterms:` (Dublin Core)
  - `rico:` (Records in Contexts)
  - `crm:` (CIDOC CRM)
  - `edm:` (Europeana Data Model)
  - `dcat:` (DCAT-AP)
  - and others used in European data spaces.
- For **industry standards** (like ISWC, ISRC, and DDEX) that are not published in formal RDF/OWL form, the Observatory provides a minimal ontological scaffolding under the `Open Music Ontology (omo)` namespace. The source files of this minimal ontology are held at https://github.com/dataobservatory-eu/open-muisc-ontology

This ensures that:

- industry identifiers can be represented and queried alongside heritage data,
- and semantic equivalence can be maintained across research, rights, and cultural collections.

In short, **OMO acts as a semantic bridge**, not as a new ontology.
In summary: The Open Music Observatory's semantic stack *connects* ontologies — it doesn't reinvent them. It provides a thin, open, interoperable layer across public, research, and industry vocabularies to keep European music data FAIR and reusable.

In several crucial domains — particularly DDEX release modelling, contributor roles, and industry identifiers — no official ontological form exists, even though these standards

make clear *ontological commitments*. As part of the Open Music Observatory's coordination workflow, we therefore export these commitments into machine-readable, future-proof RDF/OWL form, ensuring that concepts defined only in prose or PDFs can participate in a data sharing space. This work is carried out in the *Open Music Ontology* (OMO), a minimal bridging vocabulary that expresses the semantics of key industry terms (e.g. ISWC, ISRC, release types, contributor roles) without redefining or replacing the authoritative standards. OMO provides neutral, interoperable helper classes and properties — such as `omo:Recording`, `omo:MusicalWork`, `omo:ReleaseCategory`, `omo:hasISRC`, and `omo:hasContributorRole` — that allow DDEX-aligned metadata to coexist with GLAM, research, and open-data vocabularies. This ensures that industry metadata can be queried alongside cultural heritage data, and that future systems can rely on stable RDF/OWL definitions rather than informal textual descriptions (Antal 2025e).

In a data sharing space, the shared "data model" cannot be a single unified ontology. Instead, a DSS must implement a semantic mediation layer capable of interpreting and mapping *heterogeneous*, *co-existing* conceptual models.

This is consistent with the Data Space Blueprint, the European Interoperability Framework (EIF), and Gaia-X design principles, all of which explicitly require *semantic interoperability* rather than schema unification.

Our approach implements this through a **Wikibase-based mapping layer** using:

– *equivalence relations (owl:sameAs / near-equivalents)*, – *pattern-based re-expression of relationships*, and – *lightweight ontological fragments* (ontology design patterns).

This allows the data space to reuse established ontologies (CIDOC CRM, DCTERMS, RiC-O, DDEX, Music Ontology, Polifonia ODPs) while **avoiding ontology hijacking** and respecting the epistemic structures of domain-specific data (e.g., music, Finno-Ugric heritage).

Technically, this produces a *federated semantic layer*, not a single global schema — exactly the architecture envisaged by BDVA and Gaia-X for cross-domain data spaces.

5.1.1 Lightweight Ontology Patterns

We emphasize modular ontological patterns over rigid global alignment with identifying reusable semantic fragments.

```
### Classes
:Agent            a owl:Class .        # aligns with prov:Agent, crm:E21
:Work             a owl:Class .        # aligns with mo:MusicalWork, dct:BibliographicReso
:Role             a owl:Class .        # internal role concept

### Properties
:hasRole          a owl:ObjectProperty .
:roleOf           a owl:ObjectProperty ; owl:inverseOf :hasRole .
```

```
:roleType         a owl:DatatypeProperty .

### Instances
:Work123 a :Work .

:Person456 a :Agent .

:Contribution789 a :Role ;
    :roleType "composer" ;
    prov:wasAssociatedWith :Person456 ;
    dct:subject :Work123 .
```

```
:Work123 wdt:P86 :Person456 .            # "composer" (WM property)

# But we want *more semantic detail*, so we reify:
:Work123 p:P86 _:stmt1 .

_:stmt1 ps:P86 :Person456 ;
        pq:P453  :RoleComposer ;         # fake property: role type
        prov:wasDerivedFrom :SourceXYZ .
```

Where:

- `P86` = "composer"
- `P453` = "role type" (made-up but plausible)
- provenance is given by `prov:wasDerivedFrom`

This fragment is lightweight but powerful because:

- It does not force the import of full CIDOC-CRM / MO / PROV fully; which would invite contradictions among models designed for different use cases.

5.2 Multiple roles, multiple workflows to support

The need to model multiple contributions per person does not arise only from semantic complexity; it arises equally from the organisational complexity of the music and cultural heritage ecosystem. The European Interoperability Framework (EIF) makes this explicit. It emphasises that interoperability is not merely a matter of exchanging data, but of aligning legal, organisational, semantic, and technical layers so that institutions with different mandates can actually work together in practice.

In our domain, these organisational differences are profound. A single music professional—a composer, lyricist, performer, teacher, field collector, journalist, or scholar—appears in the

workflows of multiple institutions, each operating under different constraints, professional cultures, and data traditions.

Rights managers (CMOs) track authorship, contractual shares, mandated roles, local membership, and payment obligations; their systems are built around ISWC, ISRC, DDEX message flows, and national royalty rules.

Libraries describe publications, field notes, monographs, pamphlets, or educational material created by the same person, expressed via DCTERMS, MARC, and authority files such as VIAF or national bibliographies.

Archives preserve and describe primary research and documentation—field recordings, notebooks, interviews—using Records in Contexts (RiC-O) or legacy finding aids, with an emphasis on provenance rather than artistic authorship.

Music distributors and digital platforms focus on releases, rights clearance, credits, platform metadata, and consumption metrics—very different operational purposes again.

All these perspectives refer to the same individuals, yet each institution's workflow interprets the person differently. For a rights manager, the same person is primarily a rights holder or contractual party; for a library, an author or editor; for an archive, a creator of records or a subject of description; for a music service, a performing artist or contributor. These classifications are valid within their organisational contexts, but not easily reconciled into a single conceptual model.

This is where the organisational layer of the EIF becomes essential. It recognises that organisations have different goals, responsibilities, mandates, and lifecycle workflows for working with the same entities. Because these workflows are not uniform, the underlying data models cannot be forced into uniformity either. To do so would impose one organisation's worldview onto another, which the EIF explicitly warns against.

Therefore, in a data sharing space we deliberately avoid semantic conformity and instead pursue semantic interoperability:

- we do not require a single ontology;
- we do not impose unified classes or roles on all participants;
- and we do not collapse diverse workflows into one abstract conceptualisation.

Instead, the data sharing space provides a semantic mediation layer that allows, for example, a rights-management IT system and a library application—or an archival suite and a music distribution platform—to interoperate even though their internal conceptualisations are incompatible.

This is precisely why we adopt modular, pattern-based representations of roles and contributions. They allow a single person to have any number of contributions across institutional contexts without forcing libraries to adopt rights-management vocabulary, without forcing archives to adopt music ontologies, and without requiring commercial metadata pipelines to express provenance in the style of libraries or museums.

In other words:

the data sharing space preserves organisational autonomy while enabling semantic connections.

The goal of the data sharing space is not data exchange for its own sake. The goal is to let heterogeneous systems work together across the lifecycles of cultural material, even when they treat the same agents and works very differently. Librarians, archivists, rights managers, ethnographers, cultural researchers, and music distributors all produce and need information about the same individuals—but they do so for different purposes, at different times, under different constraints, and using different data models.

The data sharing space respects these differences rather than erasing them. Its contribution is to connect, not homogenise. By allowing a person to accumulate multiple, context-specific contributions—and by permitting these contributions to be classified differently in different organisational settings—the DSS creates an interoperative ecosystem without imposing a universal conceptual schema.

This approach is completely aligned with the EIF:

- *Legal interoperability*: each institution retains its own rights and contractual frameworks.
- *Organisational interoperability*: each institution keeps its workflow and functional logic.
- *Semantic interoperability*: meanings can be mapped without requiring conformity.
- *Technical interoperability*: the systems exchange identifiers, links, and reference patterns.

This is why a data sharing space succeeds where unified ontology engineering fails. It supports real institutional work, not an abstract harmonisation of concepts, enabling rights management systems, library catalogues, archival suites, and music distribution platforms to collaborate while remaining true to their respective missions.

5.2.1 Polyhierarchy

The issue of *polyhierarchy* is not a minor edge case: it is at the heart of why large-scale, multi-domain knowledge graphs—such as Wikidata, or our own data sharing spaces—struggle with representing cultural and musical reality accurately. In the *Wikidata Ontology Cleanup Task Force* and the *Mereology Task Force*, that the data professionals of the Open Music Observatory joined, we encounter the same structural tension: different communities use the "part of" relation in incompatible ways because they operate in different hierarchical systems that reflect different organisational needs.

Wikidata currently contains probably over a hundred implicit or near-equivalent uses of the "part of" relation. Sometimes "part of" expresses a physical mereology (a page as part of a book), sometimes a conceptual hierarchy (a movement as part of a symphony), sometimes an organisational grouping (a track as part of an album), sometimes a legal–economic relationship (a track being part of a rights bundle), and sometimes an abstract

knowledge-organisation hierarchy (a concept being part of a broader field). These are not equivalent uses, and attempting to force them into a single hierarchy or a strict ontology rapidly leads to contradictions.

This is not a flaw in Wikidata—it is an unavoidable consequence of operating a general-purpose ontology used simultaneously by librarians, rights managers, musicologists, archivists, biologists, pharmaceutical researchers, geographers, and cultural heritage professionals. Each of these communities inherits its own modelling tradition. And crucially, those traditions reflect organisational workflows, not just semantic preferences.

For example, in music: - A library catalogue does not care that a CD contains ten tracks composed by different authors; it treats the CD as a loanable unit. - A rights manager cares deeply about track-level authorship because royalties and licensing depend on it. - A musicologist may care about movements and sub-movements within a single work. - A digital distributor (DDEX) distinguishes between recordings, sound files, releases, and release bundles in ways foreign to both librarians and archivists.

No single hierarchy can serve all of these purposes simultaneously.

This is precisely the problem that the archival world has grappled with for decades. In traditional archival theory, the "unit of description" could be: a *fonds*, a series, box, folder in a fonds, a file, a letter or even a page of the letter.

This is not a matter of ontological taste; it is an organisational reality shaped by the scale of collections, staff capacity, digitisation status, and cataloguing philosophy. Archives did not fail to formalise these differences because they lacked ontologists—they struggled because hierarchy itself is variable, contextual, and meaningful.

In our efforts to create a pragmatic model for the Open Music Observatory we were informed and influenced by the outcomes of the ten-year effort to produce the *Records in Contexts Conceptual Model* (RiC-CM). RiC is essentially an attempt to design an ontology that could tolerate hierarchical variability between a deep archival system of a well-staffed national archive and a small community archive. RiC does not eliminate polyhierarchy; instead, it tries to provide a flexible, graph-oriented language for representing multiple hierarchical and non-hierarchical relationships at once. The result is powerful but also demanding. It is telling that, despite the conceptual elegance of RiC, few archives have rushed to replace ISAD(G). The conceptual shift is too large, the costs too high, and the implications for organisational workflow too complex. Music archive designers will have to support ISAD(G) and RiC for decades to come. (Needless to say, the designers of RiC were fully aware of this and the gradual transition is possibly.)

Similarly, libraries never truly adopted FRBRoo, even though it theoretically offered a perfect hierarchy of Work → Expression → Manifestation → Item. Why? Because libraries operate under real organisational constraints, like existing cataloguing practices, legacy IT systems, patron-facing interfaces that users are loyal too, acquisition workflows, physical holdings and their storage,loan systems. FRBRoo's clean theoretical hierarchy simply does not map onto the messy, practical realities of the often underfunded realities of a smaller music library's operations.

This is exactly the lesson that Wikidata, and any data sharing space, must internalise: polyhierarchy is not a modelling flaw—it is a reflection of institutional diversity. Different institutions use different hierarchies not because they disagree, but because they serve different roles. A "part of" relation means something different in a conservation lab, a collective rights management organisation, a radio station's playlist editor, a digitisation workflow, or an archival fonds description.

The challenge is not to normalise all of these viewpoints into a single hierarchy. The challenge is to create a modelling environment where multiple hierarchies can coexist, where contradictions do not break the system, and where cross-domain linking is possible without forcing all contributors into one conceptual framework.

This is precisely why Wikidata is cautious about enforcing strict subclass/instance models and the the a Mereology Task Force is needed just to understand the consequences of many near-equivalent meanings of the "part of" relationship.

5.2.2 Formalisation

Advocates of data spaces sometimes describe them as "just connecting databases on an as-needed or as-permitted basis," as if connection were a loose, informal network of API calls. But this is misleading. From a legal point of view, a database is either connected or not connected: if a system can dereference identifiers, look up metadata, or reuse statements, then the obligation to respect licenses, consent frameworks, provenance, and organisational constraints is triggered regardless of how lightly the connection is described.

The same precision applies on the semantic level. Our data sharing space does not avoid formal modelling. We represent our conceptual model in the Wikibase ontology, and this can be—and in practice must be—compiled into RDF and OWL axioms so that machines can interpret the classes, properties, role patterns, qualifiers, and constraints. We use property semantics, subclass hierarchies, equivalence mappings, reified statements, and alignment patterns that are fully compatible with OWL 2 DL or OWL 2 RL, depending on the use case.

Thus, "semantic mediation without semantic conformity" does not mean informality or the absence of a schema. It means something conceptually subtler:

> The data-sharing space maintains a formally specified ontology—but one that does not require all participating institutions to conform to a single, domain-unifying conceptualisation.

The formalism exists at the mediation layer, not as a global schema that all domains must adopt.

A data space cannot function without a shared URI space, identity management, explicit class/property declarations, equivalence/near-equivalence mappings, domain/range expectations, inference patterns (even if limited), or consistent referential semantics. We just understand that we are making heavy trade offs for interoperability of systems, instead of serving one type of system's internal workflows.

5.3 Future-Proofing

Future-proofing in a Data Sharing Space means designing systems so that the knowledge we curate today will still make sense — technically, legally, and conceptually — ten, twenty, or fifty years from now. Because our audience spans librarians, archivists, music-industry professionals, rights managers, researchers, and IT staff with very different technical backgrounds, the future-proofing strategy of the Open Music Observatory must be both technically rigorous and described in familiar terms.

To make this clearer, we describe future-proofing at four levels: the data model, the technology, the semantics, and the organisational workflows.

5.3.1 Future-proofing through graph architecture

Many library and industry IT systems are still based on relational databases (MySQL, Oracle, SQL Server) or on simple tables (Excel, Access). Increasingly, people have heard the word "NoSQL," even if they haven't used such systems directly, and associate it with flexibility and modernity.

A knowledge graph is a type of NoSQL system — but it is more than that. In a graph database:

- the schema is not a separate document living in an IT department's folder;
- the schema is encoded in the data itself;
- every relationship (composer of, recorded at, part of, published by) is explicitly stored alongside the entities.

This means:

- the structure of the data can evolve without breaking old records;
- old data remains meaningful even after conceptual models change;
- future systems can read the RDF graph and rebuild the entire database without needing to know the original software.

Sadly, we often hear about cases when a developers passed away, or retired, and the latest database schema only existed in their heads. Graph databases connect the schema definition to every "table cell" forever. Often, our metadata repair work is re-discovering the not formalised schema of a legacy system.

For underfunded library IT environments, this is crucial. A "TTL dump" or "JSON-LD export" is not just a backup — it is the guarantee of a future database, often with a path to transitioning to a cheaper open-source loan or archive management software with the explicit help of documenting the data first. In a music label, such future proofing of Excel repertoires helps to automate catalogue transfer to a more affordable or better quality distributor.

5.3.2 Stabilising definitions through internationally defined standard vocabularies

When libraries, archives, or rights organisations use an SQL database, the meaning of a field often lives only inside that software. For example:

- "CreatorName" inside an old PHP/MySQL-based webshop might mean a composer, or a performer, or someone who once clicked the upload button.
- An "Author" field in a legacy library system may mix lyricists, arrangers, field collectors, annotators, and editors.

But DCTERMS, RDFS/OWL, and DDEX give clear, internationally defined meanings to these concepts.

This means:

- once your data is mapped to these standard vocabularies,
- future systems — even ones not invented yet — will know how to interpret it.

This is why DDEX is so powerful: even if a label used a 20-year-old MySQL database, once mapped to DDEX, the meaning becomes future-proof.

The same applies to:

- ISWC (work identifiers),
- ISRC (recordings),
- VIAF/ISNI (persons),
- RiC-O (archives),
- MARC relator codes (libraries).

Even if the technology changes, the semantics remain stable.

5.3.3 Future-proofing through translatability and multiple serialisations

In legacy systems, a database backup is often useless outside its native software. For example:

- An Access `.mdb` file from 2008.
- A PHP webshop dumping `csv` files in an ad-hoc format.
- A FileMaker Pro database from a defunct project.

A graph database solves this because RDF data can be exported in many serialisations:

- Turtle (.ttl)
- JSON-LD
- RDF/XML
- N-Triples
- JSON

Each of these files is both human-readable and machine-readable, and any future graph system can rebuild the knowledge base from them — exactly the same way software can be rebuilt from source code. If you take a look at the description of the *Structural Business Statistics (OpenMuse)* dataset (Q600), it can be read as om:Q600.ttl or om:Q600.json

For cultural heritage preservation, this is vital:

> If we preserve the RDF, we preserve the meaning — not just the data.

This is why OMO's RDF exports act as both:

- a functional database backup, and
- a preservation format for future researchers, developers, and institutions.

5.3.4 Future services through institutional interoperability

Future-proofing is not only technical. It is also about making sure that libraries, archives, rights organisations, and small labels can move into the future without needing to throw away their old systems.

When a library aligns its catalogue with the SKCMDB (Slovak: hudobnadatabaza.sk) federated module of the Open Music Observatory:

- the library gains a modern semantic layer,
- the obsolete system gains a migration path,
- future library software can import the mapped RDF as its starting point.

This has already happened in Hungary and Slovakia: metadata that previously lived only inside ageing local catalogues has now been "semantic-lifted" into future-proof form.

The same applies to:

- music labels using old PHP/MySQL webshops,
- rights organisations with 1990s-era SQL systems,
- archives with hand-maintained Excel inventories,
- community collections with no database at all.

Once the data is mapped into the data sharing space, it becomes portable.

5.3.5 A concrete example: ALOADED, Livonian folk music, and DDEX

The ALOADED proof-of-concept demonstrates future-proofing in action, because it shows how archival cultural-heritage metadata, contemporary music-industry workflows, and emerging data-space standards can all meet in a single semantic pipeline.

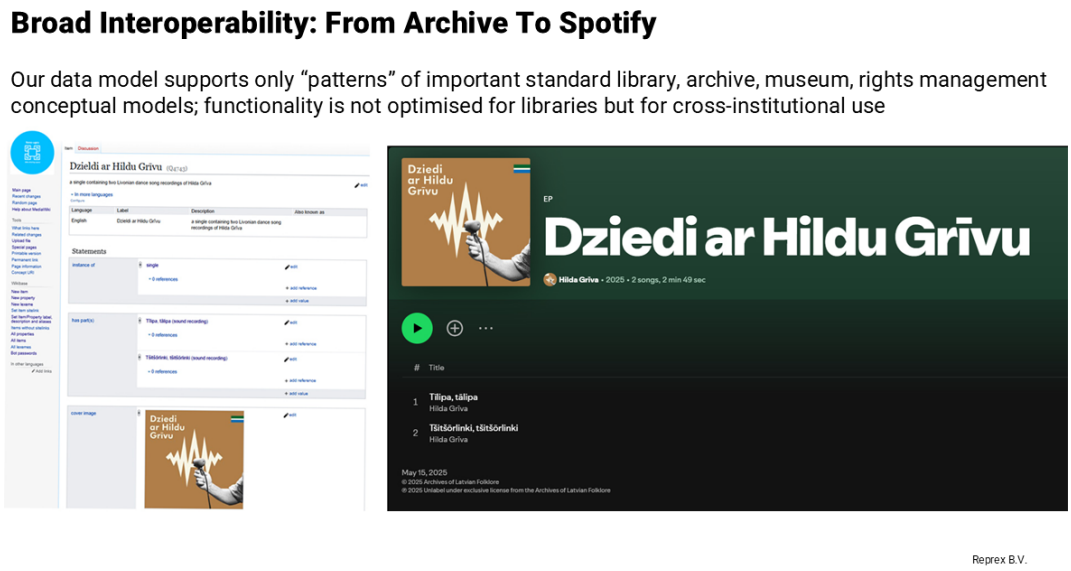


Figure 5.3: We explained this in a broader context in our *Open Access Music Dataspaces – Open Music Observatory* on LineCheck 2025. DOI: 10.5281/zenodo.17669739

Our workflow proceeded in the following steps:

- We began with archival metadata about Livonian and Latvian folk music, including the tšitšõrlinki traditional dance song Q4169, with its associated field notes, handwritten scores, and contextual ethnographic information.
- This material included a corresponding archival field recording Q4166, originally catalogued in a traditional archival system that follows a very different descriptive paradigm.
- We then expressed this material using DDEX concepts Q4745, creating mappings between archival descriptions and DDEX's contribution and release structures.

This transformation enabled several outcomes:

- the track could appear legally on Spotify https://open.spotify.com/track/184iSrExWt2K9z7S4FvJYd
- the musical score and the archival recording became findable on Garamantas.lv and Europeana
- creators and contributors could be credited transparently and correctly
- rights metadata could be clearly documented
- and the entire pipeline resulted in a FAIR, future-proof transformation of cultural heritage into a living, usable service

We describe this workflow in more detail in *A Feasibility Study With Two Datasets on the Application of the Early Stage HDTO Ontology in Practice* (Antal 2025c). There, we

explore how early HDTO (Heritage Digital Twin Ontology) structures can support the future *Europeana Data Space* and the *European Collaborative Cloud for Cultural Heritage*, including creating digital-twin representations of cultural objects that:

- have a non-commercial mechanical licence for scientific MIR processing
- and a commercial streaming-safe version (e.g., Spotify) without download or AI-training permissions

This demonstrates how one song can travel through different rights regimes and different infrastructures — all mediated through a coherent semantic layer.

This is not only a proof-of-concept for technical interoperability. It also illustrates how a shared observatory infrastructure can enable the circulation of culturally vital but economically marginal repertoires.

The Livonian ethnolinguistic community is critically endangered. Making their vocal music available in modern channels is essential for revitalisation, language learning, and community visibility. Without the shared services of the Open Music Observatory, distributing such high-cultural-value but low-market-value repertoires would be practically impossible.

A listener can now:

1. legally stream a piece of folk music (Spotify, Apple Music Classical, Deezer);
2. then borrow or download the score from a library or from Garamantas.lv;
3. a researcher can analyse the non-commercially licensed version using MIR tools;
4. and anyone can explore the archival recording, contextual notes, and documentation connected to it.

This is future-proof music librarianship — linking listening, learning, research, and cultural memory.

One of our future development tasks is to model rights, permissions, obligations, and prohibitions directly inside the Observatory knowledge graph. We outline this in *Wikibase as a Data Sharing Space: Connecting Rights, Communities, and GLAM through Federated Infrastructures* (Wikidata Conference 2025) (Antal 2025d).

6 Federated Data Modules

This chapter introduces the four federated data modules that currently constitute the backbone of the *Open Music Observatory*.

Rather than treating "data sources" as a flat list, we follow the architectural logic established in Chapter 4 and the data-to-policy pipeline in Chapter 2: data is contributed, curated, harmonised, and activated inside federated modules, each with its own governance, provenance, and semantic profile.

The *OpenMusE* project was contractually expected to populate the Observatory's four thematic pillars — *Economy*, *Diversity*, *Society*, and *Innovation* — through coordinated data collection, harmonisation, and processing. These contributions now materialise as a central *Open Music Observatory* module, complemented by three national or regional modules demonstrating how federation works in practice. Together, they show how the Observatory evolves as a network of interoperable, decentralised components rather than a single central repository.

The four federated modules presented in this chapter are:

- the **Slovak Comprehensive Music Database (SKCMDb)** — the first fully scaled national dataspace feeding the Observatory;
- the **Hungarian Music Database (HU-MDb)** — a replication and enhancement of the Slovak model, demonstrating cross-border portability and semantic continuity;
- the **Finno-Ugric Data Sharing Space** — an example of subsidiarity-based federation for regional, community, and low-depth archival collections;
- the **Open Music Observatory Core Module** — the central, pan-European module containing data created by WP1, WP2, WP3, and WP4 of the OpenMusE project, and the point of semantic linkage between national and regional modules.

Each module is described using the same structure:

1. Purpose and scope
2. Data inputs and contributing institutions
3. Metadata, identifiers, and semantic alignment
4. Governance and legal basis
5. Interoperability with OMO and other modules
6. Current status and next steps

In the sections that follow, we provide concise summaries of the existing modules and describe how they federate with the Observatory. The temporary landing page of the OMO can be reviewed on https://dataobservatory-eu.github.io/omo-landing-page/.

Once the DMP is updated, all datasets described in these modules will be linked to the DMP's Data Summaries, vocabularies, legal bases, and provenance statements, and ingestion workflows will proceed accordingly.

6.1 Slovak Comprehensive Music Database (SKCMDb)

The Slovak Comprehensive Music Database (SKCMDb) is established to make Slovak music and music-related cultural assets more visible, discoverable, and interoperable across memory institutions, rights-management ecosystems, and public collections. It is available on https://hudobnadatabaza.sk/en/.

6.1.1 Purpose and scope

- increase the precision of information about Slovak musical works, recordings, persons, and artefacts;
- improve public accessibility to sheet music, sound recordings, books, and documents;
- create a publicly accessible, linked database ("Slovak Summary Music Database") as part of a broader Slovak Music Dataspace;
- provide trusted, authoritative identifiers enabling cross-institutional coordination through VIAF, Wikidata, ISNI, library identifiers, and other persistent identifiers;
- support digital curation, enrichment, and harmonisation of metadata across libraries, archives, collective management organisations, and publishers.

This module embodies the subsidiarity principle: national institutions retain control over their data while contributing to a shared knowledge infrastructure compatible with the Open Music Observatory.

The scope of the federated Slovak module covers:

- **musical works** (compositions, arrangements)
- **recordings** (audio carriers, CDs, digitised tapes, releases)
- **persons and organisations** related to music (composers, performers, lyricists, pedagogues, musicologists, publishers, ensembles)
- **music-related documents** (books, sheet music, archival holdings)
- **physical and digital artefacts** in Slovak libraries and public collections
- **authoritative name control** for Slovak creators and entities, with linkage to international authority files
- **cross-institutional identifiers**, especially VIAF, ISNI, Wikidata Q-IDs, and internal catalogue identifiers

The scope explicitly includes both heritage-sector data (libraries, archives, SNL) and rights-management data (SOZA), forming a hybrid public-private data module.

6.1.2 Data inputs

Based on the Memorandum, the SKCMDb ingests the following categories of data:

From SOZA
- work registrations (musical works, identifiers, rightsholders) - personal data of creators and publishers (under collective-management legal basis)

additional metadata supporting identification, repertoire documentation, and cultural-value preservation

From Slovak Music Centre (MC)

- documentation of professional music creativity and live music culture across genres
- metadata from MC's internal library, databases, and central information systems
- catalogues of events, artists, performances, and publications

From Slovak National Library (SNL)

- authoritative name authority files
- VIAF IDs created for Slovak individuals and organisations not yet represented
- internal library identifiers
- catalogue records from SNL and associated Slovak libraries

From Music Fund (MF)

- publishing metadata (sheet music, books, CDs)
- catalogue and e-shop metadata (Musica Slovaca)
- information on supported creators and institutions.

From Reprex

- semantic models, schemas, reconciliation rules
- enriched metadata with global PIDs (Wikidata, ISNI, ROR, etc.)
- data linking workflows across institutions

These data inputs form a complete national module suitable for federation with regional, subnational, or thematic music datasets.

Additional libraries and memory institutions, labels, publishers, association may add: - internal identifiers - holdings metadata for music-related artefacts - distributed repertoire metadata - digitised and physical collection metadata

6.1.3 Metadata and semantic alignment

The Slovak Music Data Sharing Space adopts the same semantic modelling principles as the Open Music Observatory. Although it operates as a separate federated node with its own namespace, entity schemas, and property definitions, its conceptual model is fully aligned with OMO. All Slovak classes and properties that correspond to OMO concepts are connected through equality relations (e.g., `owl:equivalentClass`, `owl:equivalentProperty`, or the Wikidata's "equivalent to" relation), ensuring semantic interoperability across federated modules.

This alignment allows the Slovak module to use:

- the same core ontology for persons, works, recordings, organisations, events;
- the same authority-crosswalk strategy linking VIAF, ISNI, ROR, and Wikidata;
- compatible entity schemas for music works, creators, and releases;
- the same provenance and versioning rules, enabling OPA and reproducible workflows.

At the same time, the Slovak dataspace includes a number of Slovakia-specific classes and properties—for example, detailed roles in Slovak musical traditions, local institutional identifiers, historical datasets, and legacy catalogue structures.

These are maintained locally but semantically bridged to OMO through mappings. This preserves local specificity while ensuring cross-border interoperability with other national modules and with the central OMO knowledge graph.

Crucially, the Slovak module is designed for high interoperability with major open and industry ecosystems, including Wikidata, MusicBrainz, Discogs, Europeana, the Cultural Heritage Cloud, and DDEX-based commercial metadata pipelines. The SKCMDb has already exchanged a significant volume of data with Wikidata in both directions, and we are now preparing a scaled-up, systematically designed exchange. Because Wikidata is deeply integrated with MusicBrainz and Discogs, these improvements will further increase the visibility and discoverability of Slovak music across the open-music ecosystem used by curators, streaming platforms, and many other downstream stakeholders.

We are preparing our first full round-trip proof-of-concept with the WikiProject Music for December 2025. To support this, the Reprex team actively participates in the Wikidata Ontology/Cleaning Task Force and the **Wikidata Mereology Task Force**, treating this round-tripping exercise as a pilot for broader interoperability improvement across the entire music metadata landscape.

We are planning our first round-trip proof of concept with project in December 2025. For improved interoperability, we are participating in the Wikidata Ontology Cleanup Task Force and its Mereology Working Group, and treat our round-tripping excersize as a broad interoperability improvement pilot.

The SKCMDb is also the *first real-world implementation* in the music domain that operationalises several requirements independently identified by both the **CITF First Project Report (2025)** and the **Open Music Europe Green Paper (2025)**. Both documents diagnose the same structural obstacles—fragmented rights metadata, missing or unstable identifiers, incomplete provenance, and weak linkage between public cultural-heritage systems and private rights registries. The `SKCMDb` directly addresses these challenges by providing trusted identifiers, lifecycle-based provenance, cross-domain semantic mediation, and a federated governance model spanning national libraries, rights management, publishers, and public-sector institutions. This makes the Slovak module a practical demonstration of how a national music dataspace can support trustworthy AI, machine-readable copyright infrastructures, and cross-border cultural data interoperability.

6.1.4 Governance and legal basis

Governance is defined contractually in the Memorandum of Understanding and implemented through collaboration among five parties:

- Slovak Music Centre (HC) — national documentation centre for professional music culture; maintains live-music databases and coordinates IAML and IAMIC activities.
- Slovak National Library (SNL) — national authority for cataloguing and VIAF contributions; responsible for creating VIAF records for Slovak persons and organisations not yet represented.
- SOZA — collective management organisation for musical works; provides authoritative work registrations, rightsholder data, and contributes identifiers.
- Music Fund (MF) — public institution supporting music creation; contributes metadata from its publishing, catalogue, and Musica Slovaca activities.
- Reprex B.V. — data and knowledge-management provider; responsible for semantic modelling, interoperability, and technical coordination.

The Memorandum establishes a multi-party governance structure, characterised by shared stewardship over data and identifiers and coordinated authoritative control (e.g., VIAF, national library authority files, SOZA work registrations). It is legally grounded cooperation under the missions of MC, SNL, SOZA, and MF and shows openness to additional Slovak libraries and stakeholders joining the dataspace.

This positions SKCMDb as a national federated node aligned with the European Interoperability Framework and ready for integration into the Open Music Observatory.

6.1.5 Interoperability and federation

6.1.6 Status and next steps

Due to difficulties with the Data Management Planning of the OpenMusE project, the prolonged grant agreement changes, and political changes in Slovakia, building the governance model took longer than we expected. The module is being populated with data.

How SKCMDb operationalises CITF and Green Paper requirements

- **Trustworthy RMI chains** — linking SOZA work registrations with SNL authority control, VIAF/ISNI identifiers, Wikidata entities, and OMO entities.
- **Lifecycle-based provenance** — each reconciliation step includes machine-readable provenance and versioning.
- **Semantic interoperability** — crosswalks between DDEX, MARC, EDM, RiC-O, DCTERMS, Wikidata patterns, and local Slovak schemas.
- **Federated governance** — each institution retains its own data; only identifiers and mappings are shared.
- **AI readiness** — stable reference objects for works, recordings, persons, and organisations suitable for trustworthy AI guardrails and fairness testing.

Together these elements turn the Slovak dataspace into a rights-aware, culturally inclusive, and technically interoperable national module fully compatible with the Open Music Observatory.

6.2 Hungarian Music Database (HUMDb)

The *Hungarian Music Database* (`HuMDb`) is the second national-level module of the *Open Music Observatory's* federated dataspace architecture. Its purpose is to adapt and extend the principles tested in Slovakia to the specific institutional landscape, heritage depth, and rights environment of Hungary.

The conceptual basis of the Hungarian module follows the principles established in the Slovak Comprehensive Music Database (SKCMDb). As summarised in (Antal 2024a), the SKCMDb demonstrates how a national music dataspace can provide trustworthy descriptions of all music connected to a territory or cultural community, without imposing ethnomusicological or legal definitions of identity. It links public memory institutions and private rights-management organisations through a shared semantic layer, enabling legal, organisational, semantic, and technical interoperability. This framework was presented as a model for cross-border replication and for future federation with Hungarian music data owners,

showing how Hungarian institutions could join a common dataspace while retaining their own systems, authority structures, and cultural specificities. The Hungarian module applies the same principles but adapts them to Hungary's heritage depth, institutional landscape, and contemporary music workflows, forming the second national node in the Open Music Observatory's federated architecture.

At this early stage, two interoperable but distinct tracks are being developed in parallel: a heritage-focused knowledge graph in cooperation with the *Hungarian Heritage House* (HHH), and a broader music-sector knowledge base and data-sharing space developed with the *House of Music Hungary* and its library and pop music heritage collection (MZH).

we hope to include in this replication the members of the *Independent Label Fair*, i.e., Hungarian music microlabels that have a good working relationship with MZH.

These two tracks differ in governance readiness and legal clarity, but are semantically compatible, and together outline the full scope of the future Hungarian module.

6.2.1 Purpose and scope

Heritage-focused module (Hungarian Heritage House)

The cooperation with the *Hungarian Heritage House* focuses on integrating folklore, folk-music, and narrative-heritage collections into a multilingual, interoperable knowledge-graph environment. Using small pilot corpora—such as subsets of the Székely dance-music collections and the Hungarian Folk Tale Inventory—the pilot tests:

- authority reconciliation and personal-name alignment,
- geographic and settlement-level gazetteer integration,
- thesaurus and ontology alignment,
- AtoM (ICA-AtoM) round-trip export and re-ingestion.

These pilots demonstrate how legacy archival structures can be modernised and how Hungarian ethnomusicological and folklore datasets can be embedded into a wider European federation. Further documentation is provided in *Enriching and Futureproofing the Databases of the Hungarian Heritage House* (Antal and Zagyva 2025), which is available here in various formats.

6.2.1.1 Music-sector module (Magyar Zene Háza)

In parallel, the cooperation proposal with the *House of Music Hungary* (MZH) defines a richer, institution-wide music knowledge base that connects collections, studio recording workflows in line with the fixing music data at source principle[1], events,and public-facing applications into a unified data-sharing space.

[1]See the *Fixing Music Data at the Source* of our Open Music Europe Green Paper (Antal 2025f) and the European Parliament's resolution (European Parliament 2024). See Chapter 2.

The scope includes:

- harmonising library, archival, studio recording, and event workflows, and their data representation;
- creating a modern collections database interoperable with open-source library software;
- standardising legacy studio files and event data (KeleSys) through structured metadata, identifier strategies, and provenance capture;
- establishing rights-aware workflows (ISRC, ISWC, UPC/EAN, performer roles, producer rights);
- building a bilingual knowledge base capable of supporting AI-assisted interfaces, including chatbot prototypes.

The conceptual model links internal MZH systems (library, records, studio workflows, event management) to external authority systems such as VIAF, ISNI, ISRC, ISWC, Nemzeti Névtér. The knowledge base supports both internal processes (archival, rights management, acquisitions) and public-facing applications (web search, chatbot, event browsing).

6.2.2 Data inputs

Heritage module (HHH)
- folklore recordings and field notebooks - folk-tale inventory samples - dance-music metadata - archival catalogue exports - geographic and settlement-level metadata

Music-sector module (MZH)
- library and archival catalogue data - legacy collections requiring harmonisation - studio metadata and recording files - event metadata from KeleSys - collection and workflow data from web and EyeWall systems - initial rights metadata such as ISRC and ISWC

6.2.3 Metadata and semantic alignment

`HuMDb` uses the same mediation patterns as other OMO modules:

- mapping creators and organisations to VIAF, ISNI, Wikidata, and national authority files
- alignment of works and recordings with ISWC, ISRC, and DDEX categories
- mapping geographic entities through regional gazetteers, including Finno-Ugric materials
- thesaurus alignment for genres, traditions, and folk-taxonomies
- use of Wikibase schemas compatible with the OMO ontology layer

The heritage track follows semantic structures tested in Finno-Ugric pilots, while the MZH track follows contemporary rights and workflow patterns. Both converge within the same semantic layer.

6.2.4 Governance and legal basis

HuMDb governance is in formation. Current cooperation includes:

- Hungarian Heritage House: stewardship of heritage and folklore collections; cooperation agreement based on the OpenMusE grant agreement between Reprex and the HHH.
- House of Music Hungary: management of contemporary collections, studio workflows, and event metadata. ; cooperation agreement based on the OpenMusE grant agreement between Reprex and the HHH.
- Reprex B.V.: semantic modelling, identifier strategy, mediation, and future-proofing
- weCan: chatbot integration.
- optional technical coordination with KeleSys developers.

Each partner retains authority and control over its own datasets.

The emerging model follows the Slovak Memorandum of Understanding structure but adapted to Hungarian institutions.

6.2.5 Interoperability and federation

HuMDb is designed to interoperate with:

- the Slovak module, using shared schemas for persons, works, and events
- the Finno-Ugric Data Sharing Space, sharing heritage workflows and gazetteers
- Wikidata and VIAF/ISNI ecosystems
- open-source catalogue and archival systems including AtoM and Koha
- industry platforms aligned with DDEX for distribution and rights workflows

The Hungarian module introduces additional areas such as studio-file mediation and chatbot-ready knowledge-base services.

6.2.6 Status and next steps

- heritage datasets have been semantically lifted following the feasibility work (Antal and Zagyva 2025)
- the MZH cooperation document defines a two-month roadmap for studio files, events, and a public chatbot interface
- ingestion templates and minimal metadata rules are being prepared
- governance arrangements will be formalised after the first pilot integrations
- next step: establishing a unified namespace and SPARQL endpoint for federation with the Open Music Observatory.

6.3 Finno-Ugric Data Sharing Space

The **Finno-Ugric Data Sharing Space** (FUDSS) is the third federated module of the Open Music Observatory. The FUDSS can be accessed via https://finnougric.net/

It functions as a subsidiarity-based regional node, designed to support culturally endangered, minority, community, and heritage-rich datasets that lack the institutional structures available in larger countries.

This module demonstrates how the OMO federated architecture can scale to low-resource, multicultural, multilingual, and distributed memory environments, and how open-source semantic technologies allow small organisations to participate in European data spaces.

6.3.1 Purpose and scope

1. Cultural and linguistic preservation Providing a sustainable semantic infrastructure for Finno-Ugric musical traditions, including Estonian, Finnish, Sámi, Mari, Udmurt, Komi, and Livonian materials.
2. Linking heritage and contemporary music ecosystems Connecting archival field collections to rights-aware DDEX-compatible distribution workflows and multilingual knowledge graphs.
3. Demonstrating regional federation Implementing the principles described in the CITF First Project Report and the Open Music Europe Green Paper in small and distributed heritage environments.

The scope includes traditional songs, field recordings, contextual ethnographic metadata, contemporary reinterpretations, revival performances, notebook materials, archival finding aids, community-maintained materials, multilingual enrichment in Finno-Ugric languages and regional languages, and DDEX-ready metadata for selected recordings. It also includes digital twins for rights-aware distribution that separate non-commercial research uses from commercial streaming versions.

6.3.2 Data inputs

Data inputs come from four primary sources.

Latvian Archives of Folklore i.e., Garamantas.lv: - digitised field recordings and ethnographic notes - collector, performer, and informant metadata - archival structure (fonds, series, file, item) - settlement-level geographic data - Livonian and Latvian cross-border repertoires

University research datasets:
- ethnomusicology corpora - Finno-Ugric language and phonology datasets - structured vocabularies and contextual descriptions - annotations and transcriptions

Community organisations: - local archives of Sámi, Komi, Mari, Udmurt, Livonian, and other Finno-Ugric groups - recordings linked to cultural revitalisation - contextual information, translations, and performance metadata

Reprex and Unlabel: - semantic models and reconciliation rules - enriched metadata mapped to VIAF, ISNI, Wikidata, and geographic registers - DDEX catalogue-transfer metadata - digital-twin transformations

Selected distributors:
- ALOADED proof-of-concept for turning archival metadata into releasable DDEX messages

6.3.3 Metadata and semantic alignment

The Finno-Ugric module follows the same semantic principles used across the Open Music Observatory.

Key elements: - multilingual labels and scripts - reified contribution roles for collectors, performers, informants, translators - preservation of local vocabularies instead of normalisation - settlement alignment with national and European gazetteers - archival provenance aligned with Records in Contexts (RiC-O) - alignment to CIDOC CRM, DCTERMS, EDM - equivalence mappings to OMO ontology and Wikibase schemas

Industry identifiers are integrated through:
- ISWC, ISRC, and DDEX fragments defined in the OMO namespace - release metadata compatible with commercial distribution pipelines - multilingual contributor roles

This enables round-trip interoperability with archival systems, OMO, Wikidata, MusicBrainz, and commercial distributors where permitted.

6.3.4 Governance and legal basis

Governance follows a distributed subsidiarity model.

Each institution retains stewardship and decides which data to share. Rights metadata controls which versions are usable for research, public access, or distribution. Personal data follows GDPR balancing tests and cultural-heritage exemptions. Community organisations provide contextual data and approve sensitive uses. Reprex maintains the semantic layer and federated workflows.

Legal bases include public-domain status, non-commercial research licences, community authorisation, and explicitly granted commercial licences. The module implements digital-twin workflows separating versions for MIR research from versions licensed for streaming.

6.3.5 Interoperability and federation

6.3.6 Status and next steps

The Finno-Ugric Data Sharing Space demonstrates how minority and low-resource cultural communities can participate in a European data sharing space using a lightweight, federated, rights-aware approach. It connects archival heritage, community knowledge, research datasets, and modern music-industry workflows, showing how the Open Music Observatory architecture supports both cultural preservation and contemporary reuse.

6.4 Open Music Observatory Core Module

The `Open Music Europe Core Module` is the central, project-wide knowledge base that integrates the datasets, metadata structures, indicators, and workflows produced in WP1–WP4, in line with the Grant Agreement's mandate to deliver a "360-degree intelligence" system for the European music ecosystem . Its function is not to act as a monolithic database, but to serve as the semantic, methodological, and interoperability backbone that connects the thematic, national, and regional modules of the Observatory.

6.4.1 Purpose and scope

The Core Module provides the shared foundations required by the data-to-policy pipeline defined in Annex 1 of the Grant Agreement. It ensures that all datasets curated and produced within the project:

- follow the project's shared DMP, licensing, and data-protection rules
- use common indicator definitions, harmonisation schemas, and metadata standards
- expose reproducible processing workflows (survey ingestion, statistical pipelines, streaming sampling, register harmonisation)
- feed policy analysis through persistent identifiers, versioning, and provenance that fulfil the project's obligations for transparency and open policy analysis under Horizon Europe rules.

It is the reference point against which national, regional, and thematic modules align their schemas, identifiers, and provenance models.

6.4.2 Data inputs (WP1–WP4 contributions)

The Core Module ingests curated outputs from:

- WP1 policy and landscape mapping (conceptual definitions, indicator families, regulatory mappings)
- WP2 identification of data gaps, methodological foundations, and pan-European variable definitions
- WP3 data collection instruments, including survey pipelines, national statistical harmonisations, platform-data sampling, and register-based contributions
- WP4 data processing tools, ontologies, entity schemas, and automated ingestion scripts.

These curated datasets correspond to the "backend datasets" referenced in the project summary: official statistics, survey participation data, rights-holder data, and streaming-service samples that power the OMO "living policy documents"

6.4.3 Metadata and semantic alignment

The Core Module maintains the canonical Wikibase ontology layer for the project. It defines:

- shared classes for persons, organisations, works, recordings, events, economic indicators
- crosswalks between Eurostat, national statistical offices, cultural heritage vocabularies, DDEX categories, VIAF/ISNI/Wikidata identifiers
- provenance and versioning rules required for reproducible scientific workflows
- standardised definitions for contested domain concepts (composer, lyricist, average income, microlabel, participation index), in line with the project's standardisation tasks.

It acts as the semantic mediation layer between statistical, heritage, industry, and survey data—mirroring the federated architecture recommended in the CITF report.

6.4.4 Governance and legal basis

The Core Module is governed by the consortium as a whole under Annex 1 of the Grant Agreement, with SINUS as coordinator and Reprex as technical steward for semantic modelling, ingestion, and interoperability. It is also complemented by the Consortium Agreement and the elements defined in these agreements:

- The Data Management Plan of the consortium

- The OPA folders where the raw data and its contextual information, such as provenance information can be found.

All partners contribute datasets or workflows according to their WP obligations. The Module operationalises Horizon Europe requirements on open science, open-source software, transparent methodology, and reproducibility, using the licensing, privacy, and provenance conditions defined in the DMP.

6.4.5 Interoperability and federation

The Core Module is the linking hub for all other modules. It federates with:

- the Slovak Comprehensive Music Database
- the Hungarian Music Database (HHH + MZH tracks)
- the Finno-Ugric Data Sharing Space
- external infrastructures such as Wikidata, VIAF/ISNI, Europeana, the EU Open Data Portal, and the Cultural Heritage Cloud.

It does so by providing the shared namespace, identifier strategy, ontological patterns, and schema alignment that allow national and regional nodes to retain autonomy while contributing to a unified European knowledge infrastructure.

6.4.6 Status and next steps

The Core Module is operational and continuously expanded as WP3–WP4 pipelines mature. Next steps include:

- finalising the unified indicator registry
- completing automated ingestion for all WP datasets
- publishing stable RDF/JSON-LD exports for long-term reproducibility
- preparing the federation services that will onboard external stakeholders, as foreseen in the Grant Agreement and Amendment.

6.5 Summary and Integration

This modular, federated structure reflects the principles described in Chapter 2 and operationalises the architectural choices detailed in Chapter 4. The Observatory grows through interoperable contributions rather than central accumulation, aligning with the European data strategy, the Data Governance Act, and the European Interoperability Framework.

7 Data Collection

This chapter summarises how data was collected, validated, harmonised, and processed within *Open Music Europe*. It connects the practical work of WP1–WP4 to the data-to-policy pipeline described in Section 2.5 and to the governing principles defined in the Data Management Plan. It also introduces the software components developed in WP4, which implement the project's methodological requirements and prepare datasets for integration into the *Open Music Observatory*.

Because the DMP has not yet been updated to reflect all datasets collected during implementation, this chapter includes placeholders indicating where the DMP manager must add or revise content. Once the DMP is updated, this chapter will be synchronised with it and will become the operational reference for data ingestion.

> After the DMP is updated, all data-collection workflows and summaries presented here will be aligned with the approved Data Summaries and metadata structures.

⚠ Not updated

Because of the serious delay in WP6 with the updating of the Data Management Plan, this section should not be reviewed until harmonised with the DMP.

7.1 Overview of the Data-Collection Framework

Data collection in Open Music Europe followed the pipeline logic defined in Section 2.5. Each thematic work package identified indicators and conceptual models (D1.1, D2.1, D3.1), which determined the data inputs required for analysis. These sources were then evaluated based on accessibility, legal compliance, interoperability, and their relevance to indicator construction.

The data-collection effort covered four categories:

- administrative and register data
- survey data
- statistical and economic data
- platform and streaming data

The workflows for collecting, accessing, and harmonising these sources were originally documented in internal working files and the first version of the DMP. They must now be consolidated into the updated DMP.

> The DMP manager must add all final sources, access conditions, and descriptions to the Data Summaries section for traceability.

7.2 Administrative and Register Data

Administrative datasets provided essential inputs to WP1 and WP3, where economic valuation, labour structures, and participation indicators required high-resolution administrative signals. These workflows included:

- accessing royalty and licensing records from CMOs
- collecting grant and programme data from ministries
- extracting public business and organisation registers
- harmonising venue, festival, and event registers
- obtaining public-sector microdata in the Slovak pilot, including historical KULT survey data

These datasets formed sampling frames, validation structures, and reference datasets for linking surveys and platform data to legal and organisational entities.

> The DMP must describe each administrative source, its controller, its legal basis, and its access pathway.

7.3 Survey Data

Survey data in WP2 and WP3 filled gaps that administrative and platform data could not address. Survey work included:

- enterprise surveys of MSMEs in the music sector
- personal surveys on participation, wellbeing, and music behaviour
- experimental modules on diversity, mobility, and cultural citizenship
- harmonisation with national statistical surveys, especially the Slovak KULT survey

Survey workflows required careful attention to:

- sampling-frame construction
- questionnaire metadata
- informed-consent procedures
- pseudonymisation and controlled processing
- harmonisation using SDMX, DDI, and GSIM concepts

> The DMP must contain questionnaire versions, variable lists, consent protocols, and pseudonymisation details for each survey.

7.4 Statistical and Economic Data

Statistical sources enabled European and national comparability across indicators. These included:

- national accounts and satellite cultural accounts
- labour-force data
- business demography and structural-business statistics
- external-trade and export data
- cultural consumption and household budget surveys
- price indices and cost-structure information

These datasets were central to the economic modelling in WP1, circulation and diversity indicators in WP2, and societal-impact work in WP3.

> The DMP must include references, licences, and access conditions for all statistical datasets used.

7.5 Platform and Streaming Data

Platform datasets (WP1, WP2, WP4) were collected via API-based sampling and automated scripts, including:

- Spotify API samples (popularity, metadata, audio features)
- YouTube API samples
- playlist-localisation datasets
- automated crawlers for repertoire discovery

These data supported: - digital-market structure analysis - validation of economic indicators - repertoire and rights linkage - circulation and localisation studies

Licensing and terms-of-service constraints mean these datasets have strict reuse limitations.

> The DMP must describe permitted uses and restrictions for each platform dataset.

7.6 Processing and Harmonisation

After collection, datasets underwent several processing stages aligned with the architecture described in Chapter 4.

These stages included:

- cleaning and transformation
- pseudonymisation of personal data

- cross-linking with authority files and identifiers
- metadata enrichment
- structural harmonisation (SDMX, DataCite, DDI)
- conversion to formats suitable for ingestion into the OMO

The **openmusic-pipeline** (WP4) implemented these processes in R, ensuring reproducibility and alignment with FAIR and OPA principles.

> The updated DMP must list controlled vocabularies, classifications, and harmonisation rules used.

7.7 Software Components Developed in WP4

WP4 created a suite of open-source tools that implement the pipeline and prepare data for the OMO. These are described here briefly, with technical detail deferred to annexes.

7.7.1 Data-Ingestion Tools

These tools handle imports from Excel, CSV, SQL exports, APIs, and legacy systems. They include: - connectors for CMO and/or ministry datasets - survey-import scripts - API wrappers for streaming platforms

7.7.2 Validation and Reconciliation Tools

These tools perform semantic alignment and quality checks: - authority-control reconciliation (ISNI, VIAF, ORCID, corporate registries) - SPARQL-based constraint checks - duplicate detection and entity merging tools

7.7.3 Harmonisation and Metadata Tools

These implement the project's semantic rules:

- SDMX structure builders
- DDI variable metadata generators
- DataCite dataset metadata templates
- vocabulary management utilities

7.7.4 OMO Integration Tools

These tools prepare processed data for the Observatory:

- Wikibase ingestion scripts
- URI stabilisation and PID mapping utilities
- JSON-LD and RDF exporters
- R and Python client libraries for the OMO API

The DMP must reference the software components used for metadata production and processing, as required by Horizon Europe guidelines.

7.8 Position of Data Collection and Processing in the Pipeline

Data collection and processing bridge the conceptual work of WP1–WP3 and the semantic and technical infrastructure described in Chapter 4. They provide:

- validated raw inputs
- harmonised metadata
- cross-domain entity linking
- analysis-ready datasets

These steps ensure the construction of traceable, reproducible indicators and policy outputs.

7.9 Integration with the Open Music Observatory

Once the DMP is updated and the Data Summaries approved, all data collections and processing workflows described here will be integrated into the OMO:

- datasets will be linked to persistent identifiers
- metadata will be converted to human- and machine-readable forms
- provenance will be documented in accordance with OPA and FAIR
- ingestion routines will run automatically using WP4 tools

Ingestion will begin once the DMP contains the complete, updated Data Summaries and metadata structures for all datasets.

8 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.

> ⚠ Not updated
>
> This section was created at an early planning stage, and had not yet been updated. This is no longer applicable, and should not be read or quoted.

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?

8.1 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.

Some elements of DDI are described with RDF: The DDI-RDF Discovery Vocabulary is a draft specification of the DDI Alliance. (Hartmann et al. 2024). Whenever possible, we rely in our observatory with this annotation; if that is not yet possible, we follow the *DDI Lifecycle (3.3) Documentation* (Data Documentation Initiative 2020).

8.2 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, De 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*.

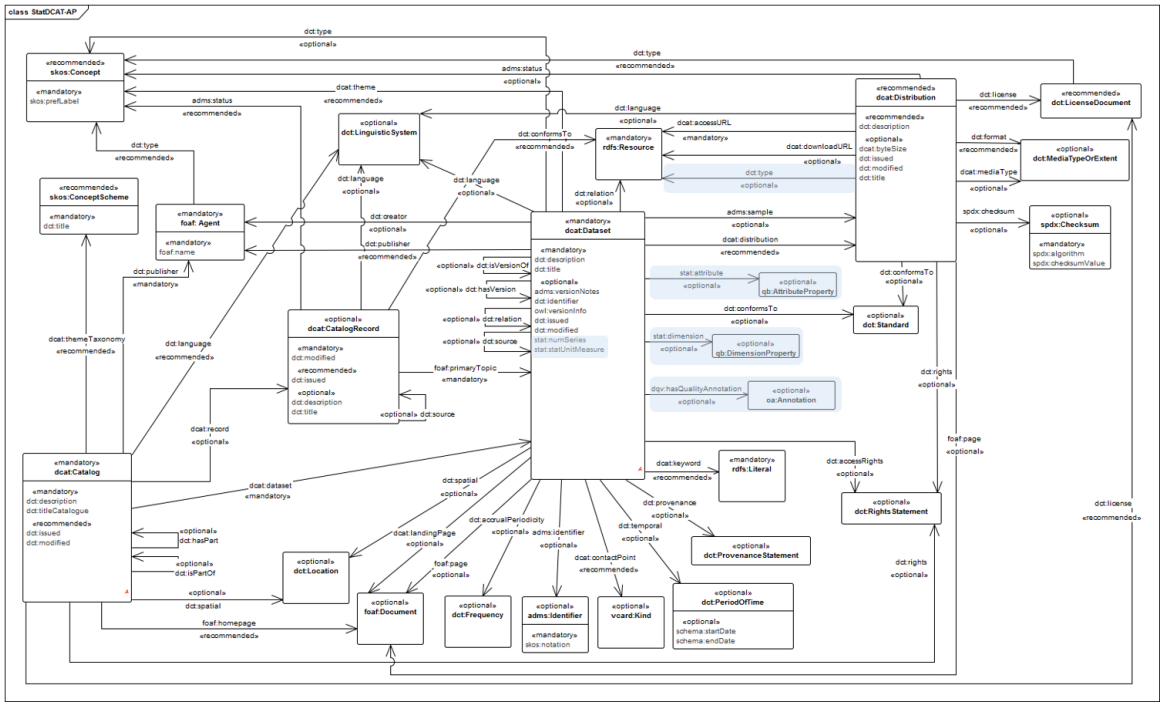


Figure 8.1: Our most important reference is the DCAT-AP 3.0 specification, and its extension to statistical data by the EU Open Data Portal.

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 (Europeana 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 (International Council on Archives Expert Group on Archival Description 2023; Hartmann et al. 2024).

> **Note**
>
> 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.

8.3 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.

> **Tip**
>
> 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).

> **Tip**
>
> 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 9.3).

8.3.1 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.

8.3.2 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 2017c)

8.3.3 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 2019b; International ISRC Registration 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 2017c)

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.

8.3.4 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.

8.3.5 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.

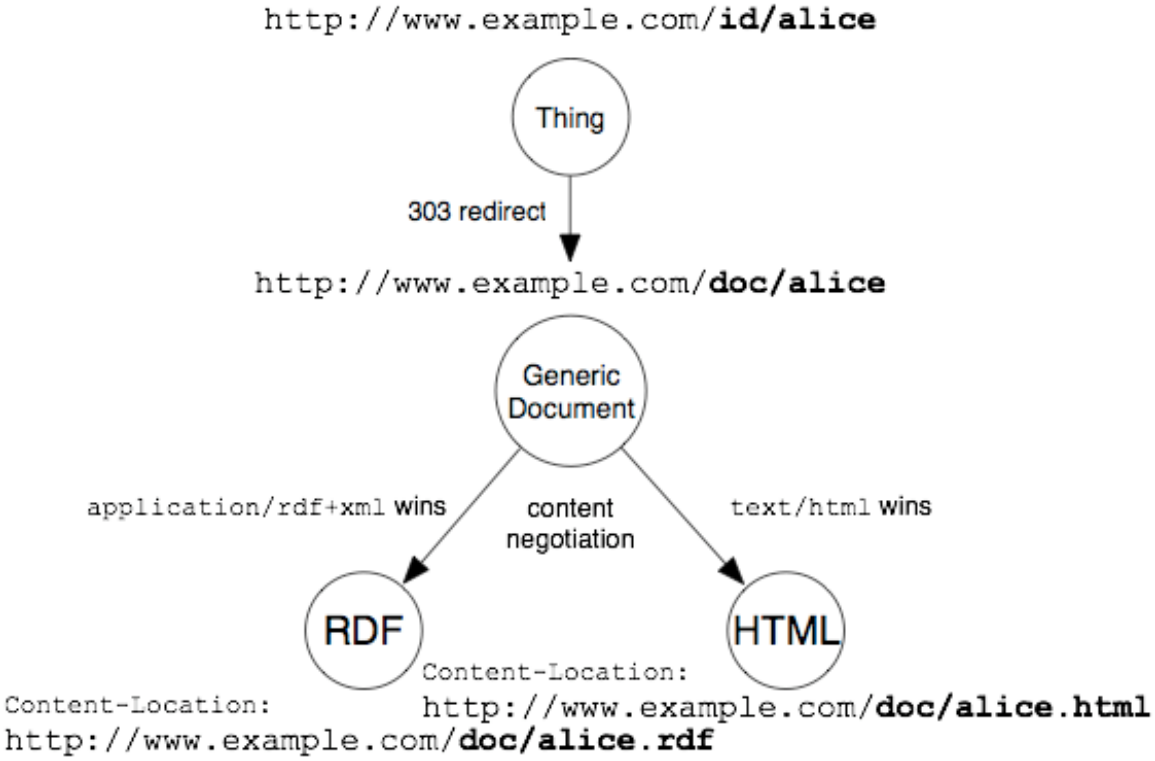
ISCC registration becomes necessary when an ISCC code needs to be globally unique, publicly discoverable, resolvable, owned or authenticated. While these features inevitably require some kind of registry, not all of them require a centralised institutional registry. The ISCC specifies the necessary protocols 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.

8.3.6 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.

9 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.

> ⚠ Not updated
>
> This section was created at an early planning stage, and had not yet been updated. This is no longer applicable, and should not be read or quoted.
> However, this part is clearly articulated in our green paper's Fixing Music Data at the Source chapter (Antal 2025f).

9.1 Value-Added Data Services

9.1.1 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.

Data sharing can reduce the redundancy in costly metadata collection, improvement, linking, updating activities, which are currently done often without coordination paralell among various authoritative database managers (for example, VIAF for libraries, ISNI, and ISRC or ISWC.) Data sharing can also greatly reduce the redundancy of parallel work at the level of collection managers, like individual libraries, publishers, labels, collective management organisations.

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.

9.1.2 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.

The fix-the-data service can mean replacing missing our outdated metadata (such as a name change of a natural person or a corporate body), or recalculating aggregated accounting or statistical data after base change, or forecasting data that is not yet available.

> Tip
>
> 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.

9.1.3 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.

> Tip
>
> 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.

9.1.4 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** is a document [in modern usage, usually a database], in which data are entered in a formal manner by a statutory authority (ISO 2017b). In statistical data collection 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.

> **Note**
>
> 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.

9.2 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.

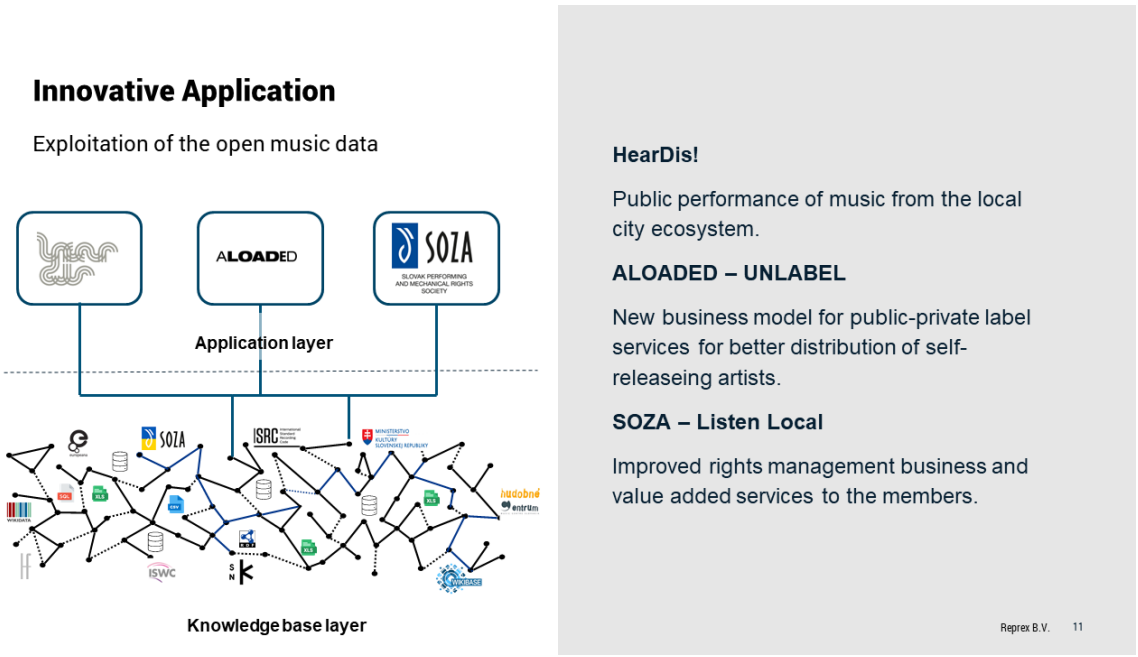


Figure 9.1: app

9.2.1 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*.

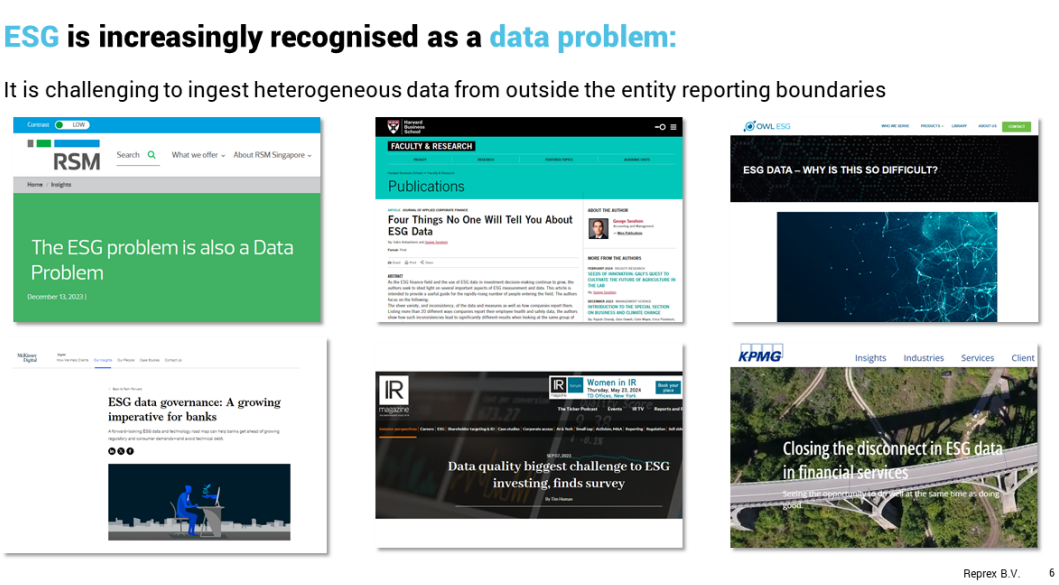
Modern platforms, such as Spotify, use similar identifiers. For example, the Spotify Artist ID `2fIUlieTjLTaNQUIKHX5B8` resolves to Celeste Buckingham's available recordings on the platform via the URL https://open.spotify.com/artist/2fIUlieTjLTaNQUIKHX5B8.

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.

9.2.2 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.

9.2.3 Listen Local

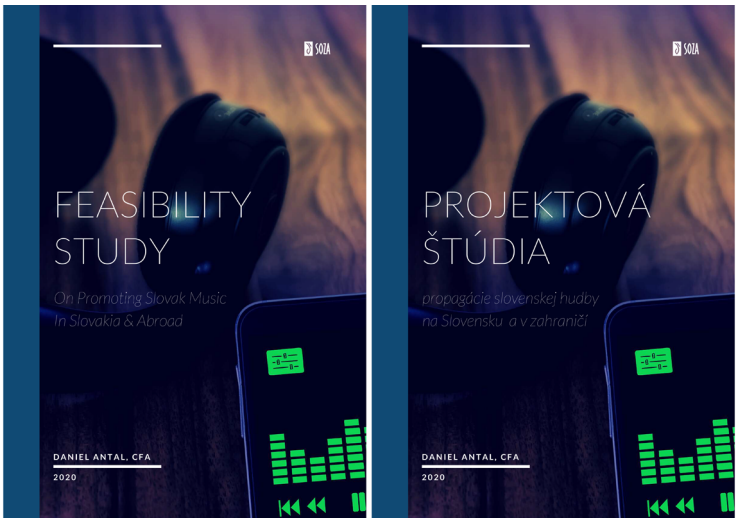


Figure 9.2: The Feasibility Study On Promoting Slovak Music in Slovakia And Abroad is an important background of our project.

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.

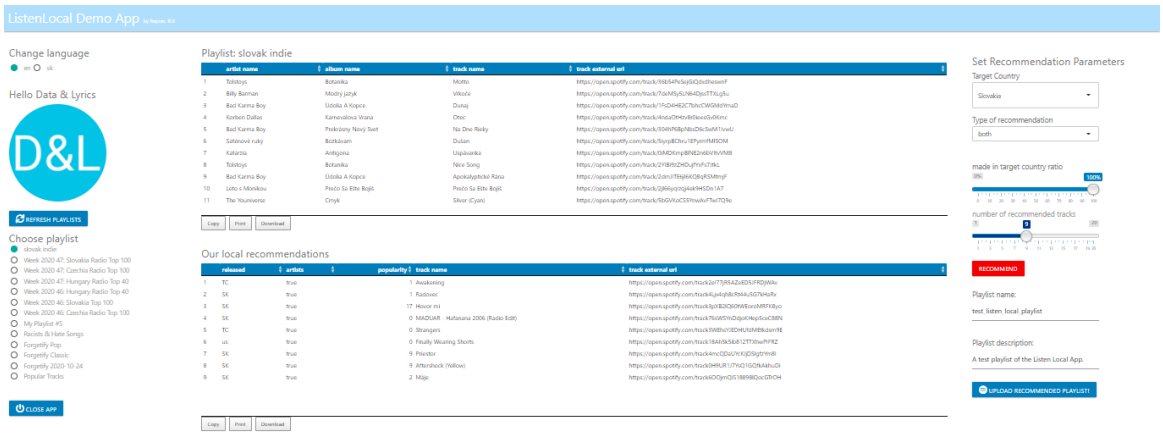


Figure 9.3: Our conceptual demo application was able to make recommendations on voluntarily meeting the local content guidelines, but it was only supported by a relatively small Slovak Demo Music Database, and could only work with Spotify, which has the most transparent and open API of all streaming 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.

9.2.4 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.

9.3 Use of AI systems

For the entity linking, related to our planned value added Section 9.2.1, 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.

10 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 10.2). The data curators of the Observatory Stakeholder Network (see Annex) 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 10.1).

Not updated

> **Warning**
>
> This section was created at an early planning stage, and had not yet been updated. Unfortunately, due to the problems of the WP6 Data Management Plan task we cannot yet show how we will fill up the pillars of the observatory.

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.

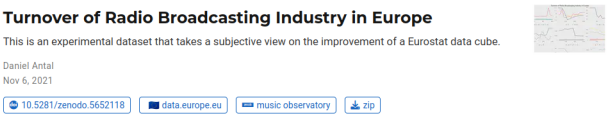


Figure 10.1: The music.dataobservatory.eu/tag/music-economy/ URL lists the downloadable datasets on the Open Music Observatory website. They can be found on EU Open Data Portal, too.

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.

10.1 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.

10.2 Topical Pillars

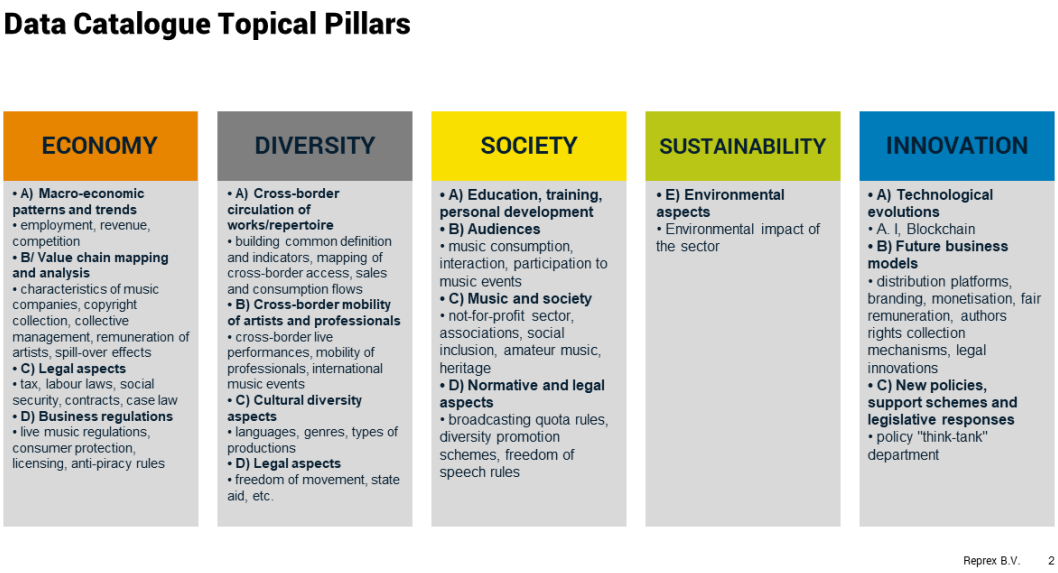


Figure 10.2: The extended five pillars, with sustainability added.

Note

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).

10.2.1 Music Economy

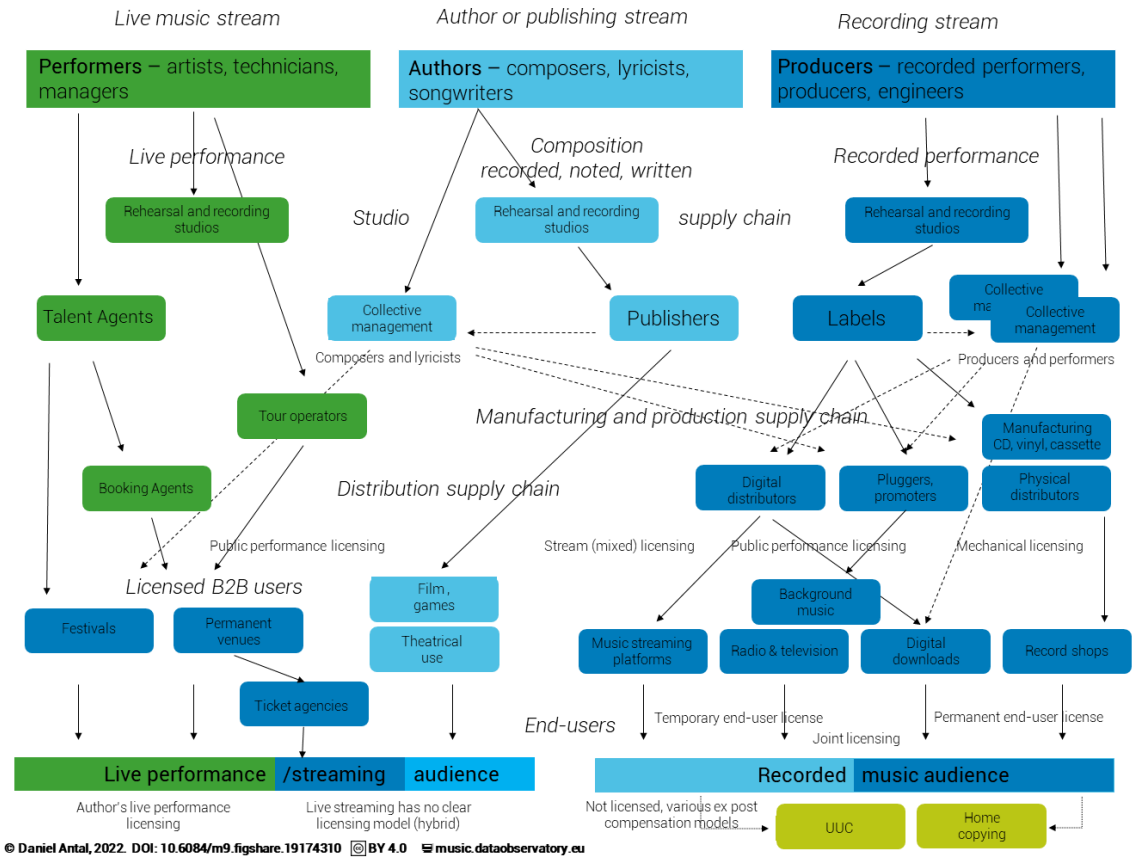
Note

Main potential data-collection and research areas identified at this stage: Macroeconomic 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.

Figure 10.3: 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.

10.2.2 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 a given 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.)

> **Note**
>
> 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)

10.2.3 Music Society

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.

> **Note**
>
> 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)

10.2.4 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 market place, 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.

10.2.5 Sustainability

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 pillar 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 9.2.2.

References

2020, SEMIC. 2020. "Providing Sustainable Data Services Through Wikibase and Wikidata." https://joinup.ec.europa.eu/sites/default/files/custom-page/attachment/2020-11/Parallel-track-4_B-Fischer_J-Thill_A-Angjeli%20final%20ppt.pdf.

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.

———. 2024a. "A szlovák adatkicserélési tér magyarországi föderációjának lehetőségei." In *Az oktatás, a kutatás és a közgyűjtemények digitális transzformációja felsőfokon : NETWORKSHOP 2024 : 33. Országos Informatikai Konferencia : 2024. április 3–5. Eszterházy Károly Katolikus Egyetem, Eger*, edited by József Tick, Károly Kokas, and András Holl, 192–98. Budapest: HUNGARNET Egyesület. https://doi.org/10.31915/NWS.2024.25.

———. 2024b. "Building a Music Data Sharing Space with Wikibase." Open Music Observatory. https://doi.org/10.5281/zenodo.17078911.

———. 2024c. "Trustworthy AI and Data-Sharing Spaces for the Slovak Music Centre. Poster Presentation at the IAMIC Conference 2024 on November 21, 2024, at Music Austria, Vienna." Open Music Observatory. https://doi.org/10.5281/zenodo.16540605.

———. 2025a. "SKCMDb. Interoperability of Music Libraries and Archives with Public and Private Music Services. Presentation at the IAML 2025 Conference in Salzburg, Austria Held on the 7th of July 2025." Open Music Observatory. https://doi.org/10.5281/zenodo.16634558.

———. 2025b. "Slovak Music Data Sharing Space. Poster Presentation at the IAML 2025 Conference in Salzburg, Austria Held on the 7th of July 2025." Open Music Observatory. https://doi.org/10.5281/zenodo.15814286.

———. 2025c. "A Feasibility Study with Two Datasets on the Application of the Early Stage HDTO Ontology in Practice." https://doi.org/10.5281/zenodo.17492178.

———. 2025d. “Wikibase as a Data Sharing Space: Connecting Rights, Communities, and GLAM Through Federated Infrastructures. Presentation on Wikidata Conf 2025.” Open Music Observatory. https://doi.org/10.5281/zenodo.17496740.

———. 2025e. “Open Music Ontology.” Open Music Observatory. https://doi.org/10.5281/zenodo.17541357.

———. 2025f. “A Green Paper on AI, Data Governance, and Metadata Policies for Europe’s Music Ecosystem.” Open Music Observatory. https://doi.org/10.5281/zenodo.17244314.

Antal, Daniel, Mester Anna Márta, Ieva Pigozne, and Mihaly Nagy. 2025. “Dataset of the Multilingual Gazetteer of the Settlements on the Livonian Coast of Northern Kurzeme.” Finno-Ugric Data Sharing Space. https://doi.org/10.5281/zenodo.17545241.

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.

Antal, Daniel, and Anna Márta Mester. 2025. *Open Music Registers*. https://doi.org/10.5281/zenodo.14767717.

Antal, Daniel, Ieva Pigozne, and Anna Márta Mester. 2025. “Remapping the Livonian Coast: A Multilingual Gazetteer of the Settlements of Northern Kurzeme.” Finno-Ugric Data Sharing Space. https://doi.org/10.5281/zenodo.15668712.

Antal, Daniel, and Natália Zagyva. 2025. “Enriching and Futureproofing the Databases of the Hungarian Heritage House.” Open Music Observatory. https://doi.org/10.5281/zenodo.17759777.

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.”

BDVA/DAIRO. 2023. “Data Sharing Spaces and Interoperability: BDVA Discussion Paper.” BDVA/DAIRO. https://www.bdva.eu.

BDVA/DAIRO Federation Working Group. 2023. “Federated Data Spaces: Position Paper.” BDVA/DAIRO. https://www.bdva.eu.

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.

Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. “Beyond VIAF Wikidata as a Complementary Tool for Authority Control in Libraries.” *Information Technology and Libraries* 40 (2). https://doi.org/10.6017/ital.v40i2.12959.

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.
Data Documentation Initiative. 2020. “DDI Lifecycle (3.3) Documentation.” https://ddi-lifecycle-documentation.readthedocs.io/en/latest/index.html.
Data Spaces Support Centre. 2025. “Data Spaces Blueprint V2.0 — Introduction: Key Concepts of Data Spaces.” 2025. https://dssc.eu/space/BVE2/1071251613/Introduction+-+Key+Concepts+of+Data+Spaces.
Diefenbach, Dennis, Max De Wilde, and Samantha Alipio. 2021. “Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph.” In *The Semantic Web – ISWC 2021*, 12922:626–42. Lecture Notes in Computer Science. Cham: Springer. https://doi.org/10.1007/978-3-030-88361-4_37.
ECHOES Ontology Task Force. 2025. “Heritage Digital Twin Ontology (HDTO) – First Draft.” Technical Report. ECHOES Project / European Collaborative Cloud for Cultural Heritage (ECCCH). https://github.com/ECHOES-ECCCH/HDTO-Heritage-Digital-Twin-Ontology.
Ernštreits, Valts. 2020. “Livonian Place Names: Documentation, Problems, and Opportunities.” *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 11 (1): 213–33. https://doi.org/10.12697/jeful.2020.11.1.09.
European Commission. 2017b. “Commission Implementing Decision (EU) 2017/1348 of 25 July 2017 on the Interoperability Framework for European Public Services (European Interoperability Framework).” *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/dec_impl/2017/1348/oj.
———. 2017a. “Commission Implementing Decision (EU) 2017/1348 of 25 July 2017 on the Interoperability Framework for European Public Services (European Interoperability Framework).” *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/dec_impl/2017/1348/oj.
———. 2020. “A European Strategy for Data.” European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020DC0066.
———. 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 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.
European Parliament. 2024. “European Parliament Resolution of 17 January 2024 on Cultural Diversity and the Conditions for Authors in the European Music Streaming Market (2023/2054(INI)).” European Parliament. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0020_EN.pdf.
Europeana. 2017. “Definition of the Europeana Data Model V5.2.8.” Europeana. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf.
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/.
Guest, Olivia, Marcela Suarez, and Iris van Rooij. 2025. "Towards Critical Artificial Intelligence Literacies." Zenodo. https://doi.org/10.5281/zenodo.17786243.
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.
International Council on Archives Expert Group on Archival Description. 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.
International ISRC Registration Authority. 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.
ISO. 2012. "International Standard Musical Work Code (ISNI). ISO 27729:2012." International Organization for Standardization. https://www.iso.org/standard/44292.html.
———. 2013. "ISO 17369:2013(en) Statistical Data and Metadata Exchange (SDMX)." London:United Kingdom: International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso:17369:ed-1:v1:en.
———. 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/IEC 5127:2017(en), Information and Documentation — Foundation and Vocabulary." London:United Kingdom: International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso:5127:ed-2:v1:en.
———. 2017c. "ISO 2108:2017 (En), Information and Documentation — International Standard Book Number (ISBN)." International Organization for Standardization. https://www.iso.org/standard/65483.html.
———. 2019a. "ISO/IEC 20546:2019 Information Technology — Big Data — Overview and Vocabulary." London:United Kingdom: International Standards Organisation. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:20546:ed-1:v1:en.
———. 2019b. "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.
———. 2023a. "ISO/IEC 11179-1:2023(en), Information Technology — Metadata Registries (MDR) — Part 1: Framework." London:United Kingdom: International Orga-

nization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:11179:-1:ed-4:v1:en.

———. 2023b. "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.

Kung, Antonio, Ray Walshe, and Rigo Wenning, eds. 2023. "Data Sharing Spaces and Interoperability." Big Data Value Association. https://bdva.eu/download/92/publications/3841/data-sharing-spaces-and-interoperability-bdva-discussion-paper-december-2023.pdf.

Magnus, Bart, and Olivier Van D'huynslager. 2021. "Podiumkunstendata Op Wikidata: De Stap Naar Echte Linked Open Data." Kunstenpunt / Flanders Arts Institute. February 11, 2021. https://www.kunsten.be/nu-in-de-kunsten/podiumkunstendata-op-wikidata-de-stap-naar-echte-linked-open-data/.

Mikš, Tomáš. 2025. "OpenMusE: Towards a Sustainable Licensing Market for AI Use of Protected Works." Vilnius, Lithuania: Slovak Performing; Mechanical Rights Society (SOZA); OpenMusE Consortium; Presentation at the CISAC European Committee Meeting. https://www.openmuse.eu/wp-content/uploads/2025/05/20250427_CISAC_EC_2025_OpenMusE_final.pdf.

Music Moves Europe. 2024. "Music Ecosystem 2025: Study on the Music Ecosystem." Publications Office of the European Union. Luxembourg: European Commission, Directorate-General for Education, Youth, Sport; Culture. https://doi.org/10.2766/95340.

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 Consortium. 2025. "Policy Brief: An Open, Scalable Data-to-Policy Pipeline for European Music Ecosystems." EU Horizon Europe Deliverable D5.7. Open Music Europe Consortium. https://openmuse.eu/.

Partanen, Niko, Philippe Rixhon, Karīna Bandere, Jānis Ziediņš, Pawan Kumar Dutt, Matīss Bolšteins, Matias Frosterus, et al. 2025. "Interoperable, Trustworthy, and Machine-Readable Copyright Data in the AI Era: Report of the CITF First Project." Publications of the Ministry of Education and Culture, Finland 2025:23. Helsinki: Ministry of Education; Culture, Finland; National Library of Finland; National Library of Latvia; Culture Information Systems Centre (Latvia); Tallinn University of Technology (Estonia); Valunode OÜ. https://julkaisut.valtioneuvosto.fi/.

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

———. 2015b. *Metadata*. The MIT Press Essential Knowledge Series. Cambridge, MA,

USA: MIT Press.

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.

Sardo, Lucia, and Carlo Bianchini. 2022. "Wikidata: A New Perspective Towards Universal Bibliographic Control." *JLIS.it : Italian Journal of Library and Information Science* 13 (1): 291–311. https://doi.org/10.4403/jlis.it-12725.

SEMIC Support Centre. 2023. "Wikidata and Wikibase — SEMIC Support Centre." https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/wikidata-and-wikibase.

Senftleben, Martin, Thomas Margoni, Joost Poort, Kacper Szkalej, and Etienne Valk. 2024. "Policy Brief 1: Music Metadata Mainstreaming and EU Law." EU Horizon Europe Deliverable D5.6. OpenMusE Consortium. https://www.openmuse.eu/.

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.

Stallmann, Claudia, Koen Deneckere, Ruben Verborgh, et al. 2023. "MetaBelgica Project: A Linked Data Infrastructure Between Federal Scientific Institutes in Belgium." In *Proceedings of the 19th Extended Semantic Web Conference (ESWC 2023)*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-33455-9_24.

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.

A Data Model

In order to support metadata repair of following the "Fixing Music Data at the Source" principle[1], we created a data model that aligns with library (authority) systems like Dublic Core (DCTERMS, DCMI) , the archival standard RiC[2], the museum standard CIDOC[3], the Europeana Data Model (EDM)[4] and HDTO[5] for cultural heritage dissemination, and VIAF; furthermore the DDEX[6], ISWC[7] and ISRC standards of rights management and music distribution or communication to the public; see also Section 2.3.

A.1 Releases

Most users are interested in releases or their contents. People lend releases of music from libraries, listen to releases on streaming services or buy them. If the metadata of a release is wrong, it is likely that the individual revenue earning assets, i.e., the tracks are also documented wrong.

Between 1960 and 2010 the album was the most important release format, arguably in the streaming era it is the single. It is very important that we are able correctly represent all types of releases.

For releases, the most applicable metadata vocabulary is DDEX, the music industry standard that does not exist in ontological form. As we described earlier, we placed some key DDEX expressions in ontological form for interoperability.

A.1.1 Release classes

[1] See Fixing Music Data at the Source in (Antal 2025f), and the European Parliament's resolution (European Parliament 2024).
[2] *Records in Contexts–Conceptual Model* (International Council on Archives Expert Group on Archival Description 2023)
[3] *Definition of the CIDOC Conceptual Reference Model* (Bekiari et al. 2024)
[4] *Europeana Data Model* (Europeana 2017)
[5] *Heritage Digital Twin Ontology (HDTO). First Draft.* (ECHOES Ontology Task Force 2025)
[6] DDEX is a standards setting organisation focused on the creation of digital value chain standards to make the exchange of data and information across the music industry more efficient. See <https://ddex.net/>.
[7] *International Standard Musical Work Code* (ISWC). ISO 15707:2022 (ISO 2022)

Table A.1: Release classes

Class	Description
`musical release`	The superclass for all releases that make one or more musical sound recordings available to the public or to a defined audience. In DDEX this corresponds to the Release entity, which aggregates metadata about the product and references the recordings it contains. In library and archival contexts it aligns with manifestations such as albums, singles, or other issued carriers. This class serves as the parent for all specific release types in the Open Music Observatory. om:Q232, skcmdb:Q873, fu:Q7074
`single`	A release intended to make one principal sound recording available, regardless of whether additional tracks such as B-sides, alternate versions, or remixes are included. Historically physical singles often contained two recordings (A-side and B-side), while digital singles may contain only one or may include supplementary tracks. In DDEX this corresponds to a `Release` with `ReleaseType=Single`. om:Q236, skcmdb:Q886, fu:Q4744 |
`album`	A release that presents a substantial group of sound recordings as a cohesive published unit. The number of tracks may vary, and albums may contain only a few long recordings or many short ones. The defining characteristic is that the release is issued by the rights holder or publisher as an album rather than as a single or compilation. T In DDEX this corresponds to a `Release` with `ReleaseType=Album`. om:Q230, skcmdb:Q473, fu:Q4286
`compilation`	A release that aggregates sound recordings originating from multiple sources, such as earlier releases, archival collections, or curated selections. Compilations may be issued for commercial distribution or for scholarly, archival, or thematic purposes, and they may include both previously released and newly issued recordings. In DDEX this corresponds to a `Release` with `ReleaseType=Compilation`. om:Q2028, skcmdb:Q2822, fu:Q6595

A.1.2 Release properties

The following properties connect releases as graph nodes to agents, recordings. and identifiers.

Table A.2: Release properties

Property	**Description**

`creator`	The agent or agents credited as the principal creators of the release as presented in catalogue metadata, packaging, or distributor-provided descriptive information. This corresponds to DCTERMS:creator, which is widely used in library and Europeana metadata for high-level authorship statements, and to the presentation-oriented `DisplayArtist`element in DDEX. Because such credits may combine performers, ensembles, conductors, or other roles, this property is intended only for descriptive and discovery purposes. Any precise contributor roles, rights-relevant relationships, or distinctions among performer, producer, composer, or other functions should be recorded using has associated agent with an appropriate subject has role qualifier. om:P137, skcmdb:P106, fu:P84
`performer`	An agent credited as a performer on one or more recordings included in the release. This property reflects catalogue or packaging information and may list performers who appear on different tracks throughout the release. It is intended only for high-level descriptive use. Detailed performer information for individual recordings, including specific roles or rights-relevant contributions, should be recorded using `has associated agent` with an appropriate subject `has role` qualifier at the recording level. om:P152, skcmdb:P97, fu:P408
`has associated agent`	Links an entity (such as a work, recording, release, event, or archival object) to an agent involved in its creation, performance, production, collection, or other lifecycle activities. This property is used whenever a specific contributor role needs to be recorded, and must be qualified with the subject has role property to indicate the function performed by the agent (for example composer, lyricist, arranger, producer, conductor, performer, or collector). It supports multiple agents and multiple roles and provides the detailed contributor model required for ISWC, ISRC, DDEX, and for provenance descriptions in RiC and CIDOC-CRM. om:P19, skcmdb:P117, fuds:P453
`subject has role` or `has role`	The qualifier of the property `has associated agent` which connects the role to the agent (e.g. composer, producer etc.)

`record label`	The imprint or publishing label under which the release is issued, corresponding to the Wikidata property for record label but used with greater rigour to reflect the specific label name printed on the release or supplied by the publisher or distributor. In DDEX this aligns with the `LabelName` element used to identify the imprint presenting the release. The `ISRC Handbook` recognises the releasing label as part of the contextual metadata associated with a recording or release but does not define label identity as part of the ISRC code itself. Detailed information about issuing organisations, phonogram producers, reissue labels, or catalogue ownership over time should be recorded using has associated agent with an appropriate subject has role qualifier. om:P163, skcmdb:P144
`tracklist`	Connects a release to the sound recordings it contains and establishes their order within that release. The numerical position of each recording is given using the series ordinal qualifier. This corresponds to the way DDEX represents track sequences using ResourceReference and SequenceNumber elements. The property describes the structure of a specific edition of a release and may differ across reissues or alternate versions. om:P162, skcmdb:P145, fuds:P70
`series ordinal`	A qualifier used with the tracklist property to indicate the numerical position of a recording within a particular release. This property follows the Wikidata convention for ordered membership and corresponds to the sequence number used in DDEX to define track order. It allows different editions or versions of a release to have distinct track sequences.
`has time`	A general temporal property used together with a `temporal role` qualifier to specify the type of date associated with a release. For example, a temporal role of date of release indicates the publication or distribution date of that particular edition. This structure aligns with DDEX, where release dates are typed elements such as ReleaseDate or OriginalReleaseDate, and with library and archival practice in which dates are associated with specific publication or production events. |
`location`	A general geographic property whose meaning is determined entirely by the `location applies to` qualifier's value, such as place of release, place of publication, place of manufacture, or place of reissue. This aligns with library practice, where publication and distribution places are recorded separately, and with CIDOC and RiC-O event-based modelling of publication. The base property provides the geographic link, while the qualifier specifies the role of the location in the lifecycle of the release.

`has access point`	A link or identifier that provides access to information about the release in an external system, such as a library catalogue, digital repository, webshop, or streaming-service landing page. This property supports discovery and interoperability and corresponds to access URLs commonly used in library and archival metadata. It does not express rights or licensing information.
`UPC or EAN`	A GS1 product identifier associated with a specific edition of a release, recorded as a UPC (12-digit) or EAN/GTIN-13 (13-digit) code. In DDEX this corresponds to an `ICPN` value used to identify a release in commercial distribution channels. UPC and EAN codes apply to releases rather than to individual recordings and may vary across different editions or formats of the same release. om:P161, skcmdb:P140
`Spotify album ID`	A proprietary identifier assigned by Spotify to a release available on its streaming service. This identifier applies only to releases that have been distributed to Spotify and corresponds in DDEX to a proprietary release identifier supplied by a digital service provider. It does not convey rights information and may differ across editions or market-specific versions of a release om:P107, skcmdb:P57, fu:P87.
`Discogs artist ID`	An external identifier assigned by Discogs to a specific edition or pressing of a release. Discogs identifiers may apply to commercial, historical, or archival releases and often distinguish fine-grained variations such as different pressings or formats. In DDEX terms this corresponds to a proprietary release identifier supplied by an external database. The identifier supports reconciliation and discovery but does not express rights information. om:P112, skcmdb:P115, fu:P435

A.1.2.1 Labels

A single release can be associated with several organisations over its lifetime, and these organisations can themselves change through mergers, nationalisation, privatisation, or catalogue transfers. For this reason, the record label property records only the imprint or publisher name as it appears on a specific edition of a release, following library cataloguing and discographic practice. It does not attempt to capture rights relationships or corporate history in the same field.

More detailed relationships are modelled explicitly using has associated agent together with a subject has role qualifier. The original issuing label, a later reissue label, the organisation acting as phonogram producer, and the company delivering the release to a digital service provider can all be represented in this way. These roles may differ from the imprint printed on the object, and they may vary between editions or distribution channels.

Corporate history and succession between labels or production companies are expressed as relationships between the label or company entities themselves. For example, one label may be the predecessor or successor of another, or may have been nationalised into or privatised from a larger organisation. This allows the model to represent cases such as the Tonalit catalogue being absorbed into a state-owned manufacturer and later reissued under the Hungaroton imprint, while still keeping the original Tonalit label attached to the historical release.

This separation between descriptive imprint, rights-related roles, and corporate history preserves accuracy across commercial, archival, and reissued materials and allows the same release or recording to be linked to all relevant organisations without overloading the record label property itself.

A.2 Recordings

Our top level classification is the `audio recording`. In our music observatory, we mainly store data about audio recordings that contain music. If other audio recordings, such as field notes of an ethnomusicological research or an audiobook, are added, they become part of the **publications**, not the music collection, and are described using DCTERMS/RiC/CIDOC rather than ISRC/DDEX.

For *recordings intended for distribution and communication to the public*, we follow ISRC and DDEX.

For *recordings not intended for commercial distribution or public communication*, but available for research within GLAM institutions under applicable legal permissions, we follow DCTERMS, RiC[8], CIDOC[9], EDM[10], and attach use policies or licences modelled with *ODLR* or simple licence statements.

Only recordings that fall within the scope of ISO 3901 (ISRC) and the music-focused DDEX profiles are eligible for ISRC assignment. Anything that *can* have an ISRC under these standards may have one assigned; however, *not every audio recording is eligible for an ISRC number* (for example, many archival, research-only or purely literary recordings are out of scope, even if technically sound recordings).

A.2.1 Recording classes

[8] *Records in Contexts–Conceptual Model* (International Council on Archives Expert Group on Archival Description 2023)
[9] *Definition of the CIDOC Conceptual Reference Model* (Bekiari et al. 2024)
[10] *Europeana Data Model* (Europeana 2017)

Table A.3: Recording classes

Class	Explanation
`audio recording`	Superclass for all recorded sound (musical or non-musical), including field recordings, interviews, audiobooks, and musical performances. Matches DCTERMS:Sound and EBUCore:AudioObject, and corresponds to FRBRoo concepts for recorded sound (e.g., F26 Recording). These recordings are all E73 Information Object items in CIDOC. The term is also compatible with the broad DDEX "Resource" notion but has no direct ISRC or DDEX equivalent. This class does not imply eligibility for ISRC, and is the level at which GLAM-style rights and ODLR-modeled use policies may apply to *any* audio object, including those outside the music domain. om:Q117, skcmdb:Q269, fu:Q21.
`music audio recording`	A subclass of audio recording whose primary content is musical. This class separates musical sound recordings from other kinds of audio such as audiobooks, interviews, or spoken-word field notes, which are treated under publications and described with DCTERMS, RiC, or CIDOC. This is an internal OMO semantic category used as the branching point for commercial, archival, demo, and pre-release musical recordings. It has no direct equivalent in ISRC or DDEX, although it roughly corresponds to DDEX Resources with UseType = Audio where the content is musical. om:Q118, skcmdb:Q2417, fu:Q4832.
`archival music recording`	A non-commercial musical audio recording created, collected, or deposited for archival, research, or documentary purposes. It is not intended for public communication or commercial distribution and therefore is not eligible for ISRC assignment in its archival state. Rights and access are governed by institutional use policies or licences, modelled with ODLR or similar statements. This class aligns with description practices in DCTERMS, RiC, and CIDOC-CRM rather than ISRC or DDEX. In some cases archival recordings may later be communicated to the public, but this requires a rights investigation; if eligible and intended for release, the object should be reclassified as a pre-release recording before becoming a commercial sound recording. om:Q122, skcmdb:Q2416, fu:Q4412

Class	Explanation
`sound recording (music)`	A musical sound recording that has been publicly communicated or commercially released (for example through physical media, download, streaming services, or broadcast). Recordings in this class are within the commercial domain and therefore eligible for ISRC assignment under ISO 3901, although the ISRC may have been assigned earlier at the pre-release stage. Once a pre-release recording becomes publicly available, it is reclassified as a sound recording (music). This class corresponds to the audio `Resource` used within DDEX releases and covers the recordings that form the content of published releases. Audiobooks and other non-musical spoken-word recordings are excluded. om:Q119, skcmdb:Q280, fu:Q121
`demo recording (music)`	A musical recording created for demonstration, pitching, rehearsal review, or internal artistic or production purposes, and not originally intended for public communication or commercial release. Demo recordings are not eligible for ISRC assignment while in this category and are not described using DDEX. If a demo recording later enters a commercial production workflow, it should be reclassified as a pre-release recording; if it is deposited with a GLAM institution for research or heritage preservation, it should be reclassified as an archival music recording. om:Q121
`pre-releaserecording (music)`	A recording designated for future release. This recording may or may not yet have an ISRC assigned. Once publicly released, it becomes a commercial sound recording. ISRC should be assigned (at any point) in the production workflow, even before mixing or mastering, provided the recording is uniquely identifiable. This aligns with DDEX PreRelease models and ISRC. Once it is released, it must be reclassified `sound recording (music).`

A.2.2 Alignment of archival recordings with DCTERMS, RiC, and CIDOC-CRM

Archival music recordings are part of the cultural heritage domain rather than the commercial music distribution domain. Because of this, they are best described using metadata models developed for archives, libraries, and museums. A recording such as *Tšitšōrlinki, tšitšōrlinki (archival recording)* (fu:Q4169) would be mapped as follows.

In all three ontologies, an archival music recording such as fu:Q4169 is treated as:

- an E73 InformationObject arising from a specific recording event
- held or curated within an archival or museum context
- governed by institutional or community rights

- non-commercial unless explicitly reclassified after rights clearance

These mappings allow interoperability between OMO and GLAM metadata systems without invoking ISRC or DDEX models.

DCTERMS alignment

DCTERMS or DCMI Metadata Terms is resource-centric and provides simple descriptive properties suitable for digital audio held in GLAM collections. It is mainly used in libraries and in some scientific repositories, but the Europeana Data Model is also derived from it.

Example mapping:

- dcterms:title = "Tšitšōrlinki, tšitšōrlinki (archival recording)"
- dcterms:creator = the performer or collector if known
- dcterms:contributor = other agents involved (collector, ethnomusicologist, fieldworker)
- dcterms:type = Sound
- dcterms:date = recording date
- dcterms:spatial = location of recording
- dcterms:license = ODLR or institutional access statement
- dcterms:rights = Recommended practice is to refer to a rights statement with a URI, too simple for music communicated to the public.
- dcterms:source = collection or fonds within the holding institution

This expresses descriptive, contextual, and rights metadata without implying commercial release.

RiC (Records in Contexts) alignment

Records in Context describes archival entities, their provenance, and their relationships. It replaces earlier, but widely used archival standards like ISAD(G). While most archives do not use yet RiC, RiC provides a conceptual model to translate archival metadata into a graph. Therefore we use RiC with backward compatibility possibilities to ISAD(G) and other prior archival standards.

Example mapping for fu:Q4169:

- Record Resource = the archival recording itself
- Record Set = the archival collection, fonds, or subfonds where the recording belongs
- Activity = the recording event (e.g., fieldwork session)
- Event = the event that was recorded

- Person or Group = performers, collectors, recordists
- Place or PlaceName = recording location
- Temporal Relation = various qualified event like recording date, transcription date, digitisation date.
- A rightsholder Agent can be linked with hasOrHadIntellectualPropertyRightsHolder.
- RiC-Relation:
  - was created by (fieldworker / performer)
  - was recorded at (place)
  - is part of (archival collection)
  - has rights holder (institution, community, or individual)
  - has access conditions (institutional policies)

RiC allows documenting complex provenance: who recorded it, in what context, and under what permissions.

CIDOC-CRM alignment

CIDOC-CRM models cultural heritage objects and their events.

Typical mapping for fu:Q4169:

- E73 Information Object = the content of the archival recording
- E 22 Human-Made Object or E84 Information Carrier = the physical tape or digital file
- E 7 Activity = the recording event (e.g., fieldwork session)
- E21 Person = performers, collectors, recordists
- E53 Place = recording location
- E52 Time-Span = date/time of recording
- E30 Right = rights and access conditions
- E78 Curated holding = the archival collection that holds the recording

CIDOC expresses the event-based provenance: what happened, who participated, and what physical/digital carriers exist.

A.2.3 Recording properties

The following properties connect the music recordings to releases, various agents or authoritative identifiers.

Table A.4: Recording properties

Property	Explanation
`creator`	The principal credited artist, group, or other agent named on the recording as presented in the source metadata (for example on a CD cover, LP label, or catalogue entry). In library practice this corresponds closely to DCTERMS:creator, which intentionally reflects the ambiguity of such statements. On commercial recordings this often aligns with the and `ISRC:Main Artist`[11]. or the DisplayArtist in DDEX, but the value may also reflect broader or ambiguous catalogue information, especially in classical music. Detailed contributor roles such as composer, conductor, lyricist, arranger, producer, or featured performer should be expressed using the has associated agent property with a subject has role qualifier. om:P137, skcmdb:P106, fu:P84
`performer`	An agent credited as performing on the recording, such as a vocalist, instrumentalist, conductor, or ensemble. This property expresses the basic performer relationship as it appears in catalogues, liner notes, or DDEX Resource metadata. It refers specifically to participation in the recorded performance, not to authorship of the underlying work. More detailed role information, such as soloist, conductor, featured performer, or instrumental role, should be provided using the `has associated agent` property with the `subject has role` qualifier. om:P152, skcmdb:P97, fu:P408
`has associated agent`	Links an entity (such as a work, recording, release, event, or archival object) to an agent involved in its creation, performance, production, collection, or other lifecycle activities. This property is used whenever a specific contributor role needs to be recorded, and must be qualified with the subject has role property to indicate the function performed by the agent (for example composer, lyricist, arranger, producer, conductor, performer, or collector). It supports multiple agents and multiple roles and provides the detailed contributor model required for ISWC, ISRC, DDEX, and for provenance descriptions in RiC and CIDOC-CRM. om:P19, skcmdb:P117, fuds:P453
`subject has role` or `has role`	The qualifier of the property `has associated agent` which connects the role to the agent (e.g. composer, producer etc.)

[11] See *International Standard Recording Code (ISRC) Handbook 4th Edition*, 2021, International ISRC Registration Authority, particularly Annex E. (International ISRC Registration Authority 2021)

Property	Explanation
`title`	The title of the recording as it appears in catalogue metadata or on the release from which the recording originates. This property may be recorded in multiple languages, each with its own language code, and may include translated or alternative titles where appropriate, for example in classical or traditional repertoire. It is conceptually aligned with dcterms:title, and with the `ResourceTitle` element in DDEX, although without requiring specific title types. This property refers only to the title of the recording itself, not to the title of the underlying musical work. om:P41, skcmdb:P114, fu:P391,
`duration`	The playing time of the recording, expressed as a time value (for example in seconds or using an ISO 8601 duration format). This corresponds to the duration element used in DDEX Resource metadata. Duration may be exact or, in the case of archival material, approximate where only partial information is available. om:P66, skcmdb:P62, fu:P99
`genre`	The musical genre or style associated with the recording. This may be taken from a controlled vocabulary used in commercial music metadata, from DDEX genre classifications, or from specialised archival or ethnographic genre schemes. Multiple genre values may be recorded where appropriate. This property refers to the genre of the recorded performance, which may differ from classifications applied to the underlying musical work. om:P36, skcmdb:P139, fu:P452
`language of work or name`	This property corresponds to the Wikidata property for indicating the language associated with a work or recording. In the context of sound recordings it refers specifically to the principal language used in the recorded performance, typically the language of sung or spoken content. It does not refer to the language of the recording title, which is indicated using language-tagged literals, nor to the language of the underlying musical work, which may be recorded separately at the work level. For instrumental recordings this value may be omitted, or the value “instrumental music (no language)” may be used. om:P33, skcmdb:P37, fuds:P80
`location`	A general property linking a recording to a geographic place. The specific meaning of the location is always determined by the value of the qualifier `location applies to`, such as place of recording, place of release, place of collection, thematic location, or place of creation. This avoids ambiguity between technical recording locations, descriptive subject locations, and other geographic associations. The ISRC country code must not be interpreted as a recording location, and if recorded for analytical purposes it should use its own qualifier. The base property serves as a simple place link, while all semantic detail is expressed through qualifiers.

Property	Explanation
`ISRC`	The International Standard Recording Code assigned to a musical sound recording in accordance with ISO 3901. This identifier uniquely identifies a specific recording and follows the ISRC format (country code, registrant code, year of reference, designation code). Only recordings intended for public communication or commercial exploitation are eligible to receive an ISRC. An ISRC may be assigned at any point in the production workflow once the recording is uniquely identifiable. Archival or non-commercial recordings normally do not carry an ISRC unless they are reclassified for commercial release. In DDEX metadata an ISRC is provided as a ResourceIdentifier with IdentifierType = ISRC. om:P136, skcmdb:P59, fu:P90
`appears in order on` or `appears on`	Connects a recording to a release on which it is included. When used with a series ordinal qualifier, it expresses the position of the recording within the release's track sequence. This corresponds to the way DDEX represents resources within releases, where each recording is listed as a ResourceReference with an associated sequence number. A recording may appear on multiple releases, possibly with different sequence positions. om:P156, skcmdb:P141, fu:P474
`Spotify track ID`	A proprietary identifier assigned by Spotify to a track available on its streaming service. This identifier applies only to recordings that have been distributed to Spotify and is not used for archival, demo, or other non-commercial recordings. In DDEX terms it corresponds to a proprietary identifier supplied by a digital service provider. Use of this identifier is subject to Spotify's terms of service. om:P106, skcmdb:P56, fu:P86
`Apple track ID`	A proprietary identifier assigned by Apple Music to a track available through its streaming or download services. This identifier is used only for recordings that have been distributed to Apple Music and does not appear on archival, demo, or other non-commercial recordings. In DDEX this corresponds to a proprietary identifier supplied by a digital service provider. Use of this identifier is subject to Apple's media terms and conditions.
`MusicBrainz recording ID`	An identifier assigned by the MusicBrainz database to a recording. MusicBrainz identifiers may be applied to commercial, archival, demo, or unreleased recordings, and provide a stable open identifier for interoperability with external music metadata sources. In DDEX this corresponds to a proprietary identifier supplied by an external database rather than a digital service provider. om:P109, skcmdb:P60, fu:P91

Property	Explanation
`is recording of`	Links a sound recording to the musical work that it realises or performs. This property expresses the relationship between a specific recorded performance and the underlying abstract work, whether the work is formally identified by an ISWC or represented as a traditional, anonymised, or locally defined musical work. In DDEX this corresponds to the reference from a Resource to a Work, and in ISWC–ISRC workflows it provides the essential connection between the work identifier and the recording identifier. In archival and ethnographic contexts it may link recordings to non-notated or culturally defined works. om:P122, skcmdb:P108, fu:P456

The `is recording of` property is recording of expresses the essential relationship between a particular sound recording and the musical work that it realises. In the context of ISWC and ISRC this link is fundamental, because ISWC identifies abstract musical works and ISRC identifies the recordings that embody those works, and the two systems rely on such connections for rights management, repertoire tracking, and distribution of royalties. In DDEX this relationship corresponds to the way a Resource representing a recording refers to the underlying Work in ERN and related messages, enabling digital services to associate a performance with the correct composition.

Within CIDOC-CRM the correspondence is expressed through the distinction between the conceptual musical work, treated as an E28 Conceptual Object, and the recorded performance, treated as an E73 Information Object brought into being by a performance or recording activity. CIDOC models this as an information object created through an event that actualises or realises a conceptual work, which parallels the conceptual grounding of the property in OMO/SKCMDb. RiC-O follows a similar approach, where the recording is a record resource and the work is an intellectual or cultural entity represented or instantiated by that resource, allowing the relationship to be documented even when the work has no formal composer attribution or ISWC registration.

In library and archival practice the relationship is equally important because it connects a recording to the abstract musical content that a catalogue or finding aid describes. This applies both to well-defined notated works in Western art music and to traditional, anonymous, or community-defined works in ethnographic and archival collections. The property therefore provides a consistent modelling mechanism across commercial, archival, and scholarly contexts, allowing all these systems to express the same conceptual link in ways compatible with their respective standards.

A.3 Works

The work model in the Open Music Observatory describes the abstract musical content that may be realised in performance, embodied in recordings, and made available through

releases or published scores. In this sense it follows the same conceptual boundaries used in library science, rights management, cultural heritage description, and commercial music metadata, even though these traditions express the concept in different ways. In FRBR and the Library Reference Model the work is the abstract intellectual creation that precedes any performance or manifestation. In CIDOC-CRM this corresponds to a E28 Conceptual Object that becomes realised through events that produce performances and recordings.

In the Records in Contexts ontology it aligns with an `IntellectualEntity` whose existence is distinct from the record resources that document or embody it.

In DDEX the same notion appears in the Work entity, which may be accompanied by descriptors such as titles, contributors, genre, identifiers, and proprietary references. In the ISWC standard the musical work is the registered abstract composition that connects to authorship shares, rights, and society-level administration.

The creator of a work is always an intellectual author. This may include a composer, lyricist, librettist, or another contributor who is recognised under authorship rules such as those defined by CISAC in connection with ISWC registration. The role is equivalent to the intellectual authorship sense of dcterms:creator, without implying performance. In CIDOC-CRM the creator is an actor associated with a conceptual production activity, and in RiC-O it is an agent responsible for the creation of an intellectual entity. DDEX differentiates the roles of authors and performers in the WorkDescriptor, and the ISWC standard treats these authorship roles as central to rights management. The OMO model retains this clarity by reserving creator for authorship and describing all other contributions through the has associated agent property, qualified with subject has role.

A.3.1 Work classes

Class	Explanation
`intellectual work`	The superclass for all conceptual works, including musical works. Aligns with LRM Work, CIDOC E28 Conceptual Object, and RiC-O IntellectualEntity. Used for any abstract content realised in performances or recordings.
`musical work`	A subclass of intellectual work representing a musical composition or culturally defined musical entity. This includes composed works, traditional works, anonymous folk works, and musical structures with or without notation. Aligns with the ISWC domain and the DDEX Work entity. om:Q138, skcmdb:Q270, fu:Q20
`song`	A subclass of musical work intended for vocal performance, including popular songs, art songs, folk songs, and traditional sung repertoire. Corresponds to vocal musical works in FRBR/LRM and ISWC practice. om:Q155, fu:Q23

A recording is linked to the work it realises using the is recording of property, and the work maintains the reciprocal link through has recording. This mirrors the ISWC–ISRC relationship in rights management, where a registered work may have multiple registered recordings.

DDEX models the same relationship when a Resource references the underlying Work in release metadata. FRBR/LRM treats recordings as manifestations of performances that are expressions of the work, and CIDOC-CRM distinguishes between the performed event, the recording event, and the resulting information object, all of which may be tied back to the conceptual object representing the work. In RiC-O the recording is a record resource that documents or embodies the intellectual entity. By modelling the link explicitly, OMO ensures consistency whether the context is commercial distribution, archival preservation, or library cataloguing.

Works may be discovered or accessed through published or manuscript scores, represented by has score, or through external catalogues and digital repositories via has access point. This aligns with FRBR distinctions between expression and manifestation, where a score is a manifestation of the work's notated expression, and with CIDOC-CRM where scores may be described as information carriers linked to the conceptual content they represent. In RiC-O these relationships connect the intellectual entity to the record parts or representations that document it. The model accommodates published identifiers such as ISMN and ISBN, allowing connections to library catalogues and authority files.

Identifiers play a central role in interoperability. ISWC identifies the abstract musical work under ISO 15707 and is managed within CISAC's global registration system. ISMN identifies printed music, while older scores may carry ISBNs. MusicBrainz Work IDs support integration with open linked music graphs. These identifiers allow OMO to interoperate with rights societies, libraries, archives, and public knowledge bases without altering their respective practices.

> **i** Local content and heritage
>
> The property `is heritage of` originates from the Heritage Digital Twins Ontology, where heritage attaches cultural objects to communities or cultural identities. In OMO the property is used in a dual manner that distinguishes ethnolinguistic heritage from national or regulatory heritage. Ethnolinguistic heritage refers to cultural belonging defined by language or community identity. For example, music classified as heritage of Hungarians refers to works connected to Hungarian-speaking communities regardless of national borders, and heritage of Slovaks encompasses works relevant to Slovak ethnolinguistic identity whether they originate in Slovakia, Hungary, or Serbia.
> National or regulatory heritage refers to classifications used in local content regulation, cultural policy, or territorial reporting. Heritage of Hungary includes works created in Hungary, works by minority communities within Hungary, and in some contexts works published or produced in Hungary, because national regulations treat these categories inclusively. By using qualifiers, the model can describe both the community-significance dimension and the regulatory-significance dimension without conflating them.

This alignment enables the OMO work model to operate coherently across commercial, scholarly, and heritage-oriented contexts. Commercial systems rely on ISWC and DDEX; libraries use DCTERMS, MARC, VIAF, and published identifiers; archives use RiC-O and event-based provenance; museums use CIDOC-CRM to link conceptual objects to perfor-

mances, recordings, and documentation; community heritage frameworks use classificatory properties such as is heritage of. By maintaining a clear abstract concept of the work and linking it to recordings, scores, agents, and heritage dimensions, the model allows diverse systems to interoperate without compromising the integrity of their respective traditions.

A.3.2 Work properties

Property	Explanation
`creator`	The intellectual authors of the work: composers, lyricists, librettists, or other creators recognised under ISWC/CISAC authorship rules. Aligns with the authorship meaning of dcterms:creator and the CISAC Work roles. Does not include performers or producers. Specific roles should be assigned using has associated agent + subject has role.
`has recording`	Links a musical work to the sound recordings that realise or perform it. Mirrors the Work–Resource relationship in DDEX and is parallel to ISWC–ISRC associations in rights management. om:P115, skcmdb:P107, fu:P457
`has score`	Connects the work to its notated manifestations (printed score, manuscript, digital edition). Aligns with library practice, FRBR Manifestation/Expression, and CIDOC's E84 Information Carrier for notated material.
`is curated member of`	Indicates that the work belongs to a curated collection, repertoire list, anthology, ethnographic corpus, or thematic grouping. Useful for scholarly and GLAM contexts.
`is heritage of` (local to)	Connects the work to an ethnolinguistic group, cultural heritage community, or national tradition. Supports policy use cases such as local content quotas, intangible heritage classification, etc.
`has access point`	A link to a catalogue entry, archival description, or other external resource describing the work. Mirrors the meaning of an access point at the release/recording levels without implying rights or licences.
`ISWC`	The International Standard Musical Work Code assigned according to ISO 15707. Identifies the abstract work independently of any recording or release. May be stored but not publicly displayed if required by CISAC policies.
`MusicBrainz Work ID`	A stable open identifier for the musical work in the MusicBrainz database. Useful for interoperability across open music metadata ecosystems.

A.4 Music Sheets

The music sheet classes in the *Open Music Observatory* describe notated representations of musical works, whether they are printed publications, digital editions, or manuscript

sources. In the current model these classes are treated as subclasses of a broader category of publications. This is a pragmatic choice that allows textual and notational materials to be handled within a unified interface, separate from sound recordings and audiovisual items. Although this means that manuscripts become grouped with publications, even though they were never formally published, this approach supports a coherent user experience and can later be refined without affecting the underlying semantic relationships.

A.4.1 Music sheet classes

Printed music sheets correspond closely to the concept of a manifestation in the FRBR and Library Reference Model, where the printed score or part is a physical or digital embodiment of an expression of the work. In this sense an ISMN identifies the published score and an ISBN may appear for older scores printed before the adoption of ISMN. DCTERMS can describe both printed and digital editions through properties such as title, creator, publisher, date, and identifier. Europeana practices similarly treat printed scores as text-based cultural heritage objects that carry bibliographic descriptions and link back to the abstract musical work.

Manuscript scores and handwritten sources present a different case. These items are not publications in the bibliographic sense and belong more naturally to archival or museum domains. In CIDOC-CRM a manuscript score would be an Information Carrier, generally described as an E84 Information Carrier, whose content is an E73 Information Object representing the notated music.

In RiC-O the manuscript appears as a Record Resource within a specific fonds or collection, accompanied by provenance relationships that explain who created it, when, and under what circumstances. Despite their different status, both printed and manuscript scores correspond to notated expressions of the musical work, and treating them within one superclass allows the model to present them together in a practical interface.

Table A.7: Music sheet classes

Class	Desciption
`publication`	A textual or notational resource that presents intellectual content in a fixed form. This includes printed or digital books, articles, and music scores. Manuscript materials are included for practical reasons, allowing all text and notation resources to be handled through the same interface even when they have not been formally published. Corresponds to FRBR Manifestation, CIDOC-CRM Information Carrier, and RiC-O Record Resource.
`music sheet` or `printed music sheet`	A notated representation of a musical work, either printed or digital. This includes scores, parts, and lead sheets issued as published editions. Identified by ISMN or ISBN where applicable.

Class	Desciption
`manuscript sheet`	A handwritten or otherwise unpublished notated source that documents a musical work. This includes composer manuscripts, archival copies, and field-transcribed notation. Corresponds to archival record resources in RiC-O and information carriers in CIDOC-CRM.

The relationship between a musical work and its notated representations is captured in OMO by the `has score` property. This establishes a link from the abstract work to one or more physical or digital manifestations, aligning with FRBR's distinction between work, expression, and manifestation, and with CIDOC-CRM's event and carrier-based descriptions. The same property can connect a work to a modern scholarly edition, a historical printed part, or a manuscript source found in an archive. The unified modelling of these materials allows digital publications, library holdings, and archival manuscripts to operate within the same structural framework while retaining the ability to differentiate provenance, publication status, and material form.

The classification of *music sheets* as publications therefore serves both practical and conceptual purposes. It supports the workflow in which textual and notational materials are managed separately from sound and audiovisual content, and it maintains compatibility with established bibliographic and archival ontologies.

Future refinements to the publication hierarchy can introduce a clearer separation between published scores and unpublished manuscripts without altering their fundamental relationships to works, agents, or collections. For now the model provides a coherent and interoperable way to represent all forms of notated music, ensuring that they can be linked to musical works and contextualised within the broader cultural heritage ecosystem.

A.4.2 Music sheet properties

Table A.8: Music sheet properties

Property	Description
`creator`	The intellectual author of the notated content, typically the composer or lyricist of the musical work. Does not include performers or editors. Corresponds to dcterms:creator and authorship roles in FRBR and RiC-O.
`title`	The title of the music sheet as it appears in the edition, catalogue entry, or archival description. May be recorded in multiple languages. Corresponds to dcterms:title and normal title practices in library metadata.
`has score`	Links the music sheet to the musical work that it represents. Allows multiple manifestations or manuscript sources to be associated with the same abstract work.

Property	Description
`publisher`	The entity responsible for issuing the printed or digital music sheet. Corresponds to dcterms:publisher and to publisher relationships in library cataloguing.
`has time`	The date on which the music sheet was published or first made available in its present form. For manuscript sheets this may be the date of creation if known. The full path is represented by: `has time` property with qualifier `temporal relation` and the value `date of publication`
`is curated member of`	Indicates that the work belongs to a curated collection, repertoire list, anthology, ethnographic corpus, or thematic grouping. Useful for scholarly and GLAM contexts.
`is heritage of` (local to)	Connects the work to an ethnolinguistic group, cultural heritage community, or national tradition. Supports policy use cases such as local content quotas, intangible heritage classification, for example, the ECCCH HDTO ontology.
`has access point`	A link to a catalogue entry, archival description, or other external resource describing the work. Mirrors the meaning of an access point at the release/recording levels without implying rights or licences.
`location`	The physical or digital holding location of the item when applicable, such as a library, archive, or collection. The `location applies to` qualifier specifies whether the location refers to the holding institution or the place of publication.
edition statement?	Information describing the edition, revision, or version of the music sheet. Common in printed scores to distinguish revised editions, urtext versions, or scholarly editions.
material or format?	The physical or digital format of the item, such as print, manuscript, facsimile, digital score, or part. Corresponds to material description practices in library and archival metadata and to CIDOC-CRM carrier types.
`ISMN`	Identifier for published notated music (printed scores). Applies to manifestations, but recorded here to support linking between work and published sheet resources.
`ISBN`	Identifier for older printed music publications when ISBNs were used prior to the introduction of ISMN as a distinct standard.

A.5 Agents

Agents are the people, groups, or organisations involved in the creation, performance, production, distribution, or documentation of musical works, recordings, releases, and publications. Different metadata traditions describe agents in different ways. Rights management systems such as ISRC and ISWC require precise distinctions between performers, composers,

producers, and other contributors. Library cataloguing frequently records only a creator or contributor statement based on what appears on a publication or catalogue entry, even if such statements are ambiguous. Archives may describe materials at the level of a fonds and attribute provenance to an estate, an organisation, or a collecting activity rather than to individual creative agents. The OMO model allows all three traditions to coexist by separating the class of the agent from the role the agent performs and by using role-qualified relationships when precision is required.

DDEX defines `DisplayArtist` as the credit that should be presented to users at the release level. `DisplayArtist` is a presentation and marketing field rather than a rights or authorship classification. It may represent a group of artists, a conductor and orchestra, a solo performer, or an aggregated credit such as *Various Artists*. The ISRC `Main Artist`, by contrast, is a recording-level performer designation defined in Annex E of the *ISRC Handbook* and refers specifically to the primary performing artist or group contributing to the recorded performance. A release can therefore have a `DisplayArtist` that does not match the ISRC `Main Artist` of any of its tracks.

The OMO model reflects this separation by treating creator as a general catalogue credit and reserving detailed performer, producer, or authorship information for role-qualified relationships through has associated agent.

Table A.9: Agent classes

Class	Explanation
`creator (artist)`	The main agents associated with the creation of the release, equivalent to DCTERMS:creator and DDEX:Artist. The MainArtist is a subclass of DDEX:Artist, and a mandatory field of ISRC. In case of classical music, the evolving standards clarifies the MainArtist is the performer, and the composer is a new class. Ambiguous from a copyright management point of view, and *we aim not to use this as a qualifier*, however, we connect various artists with the `creator` *property* for easier finding the music. The qualifier can be used when we do not know the role of an agent.
`composer`	Equivalent with composer (Q36834), an agent who is credited as author of a musical work
`arranger`	Equivalent with music arranger (Q1643514)
`lyricist`	Equivalent with lyricist (Q822146) who may be an author specialised in song lyrics or opera librettos, or a poem whose words are used in vocal music.
`producer`	An agent responsible for the (mastered) production of the recording. Near equivalent with record producer (Q183945) with the exception that we aim to record those producer(s) who have sound recording rights. Often several people are credited in participating in the production role but usually only the main producer has rights.

Class	Explanation
`performer`	An agent who is a credited performance on the live or recorded performance. A subclass of the DDEX:Artist class. We aim to find those performers who may be credited with performer's right; sometimes session musicians may have contributed to a performance whose role is based on work-by-hire and have no performer's right.

Table A.10: Role classes

Property	Description
`creator` alias (`main artist`)	A general class for agents credited in a catalogue, metadata feed, or publication as responsible for the creation of a release or recording. It corresponds to dcterms:creator and to the DDEX Artist class used in DisplayArtist statements. It does not imply a specific authorship or performance role. Detailed contributions should be recorded using role-qualified properties.
`composer`	An agent credited as the author of a musical work. This corresponds to composers in ISWC registration, to work authors in DDEX WorkDescriptors, and to intellectual authors in library and archival metadata.
`lyricist`	An agent credited as the author of song texts, librettos, or poems used in vocal compositions. This aligns with ISWC lyricist roles and with standard authorship roles in bibliographic metadata.
`arranger`	An agent who prepares or adapts a musical work or traditional melody for performance or publication. In rights management this may count as authorship depending on the jurisdiction and degree of creative contribution.
`producer`	An agent responsible for the production of a sound recording, particularly the entity holding neighbouring rights in European law or sound recording copyright in jurisdictions such as the United States and the United Kingdom. A single recording may list many production-related credits, but the rights-bearing producer is usually a smaller subset of these contributors.
`performer`	An agent credited as performing on the recording or release, such as a vocalist, instrumentalist, conductor, or ensemble. This property expresses the basic performer relationship as it appears in catalogues, liner notes, or DDEX Resource metadata. It refers specifically to participation in the recorded performance, not to authorship of the underlying work. More detailed role information, such as soloist, conductor, featured performer, or instrumental role, should be provided using the `has associated agent` property with the `subject has role` qualifier.

A.5.1 Roles

Roles specify the function an agent performs in relation to a work, recording, release, publication, or event. In the OMO model, roles are never stored as subclasses of agents but always as qualifiers on the has associated agent property. This allows a single agent to perform multiple roles without ambiguity and supports detailed rights management, library cataloguing, and archival provenance descriptions.

Table A.11: Role classes

Property	Explanation
`creator` alias (`main artist`)	A descriptive credit used at the release or catalogue level. It may include composers, conductors, performers, or ensembles and is not restricted to rights-relevant roles.
`has associated agent` qualified with `composer` role	Indicates that the agent is the author of the musical work.
`has associated agent` qualified with `lyricist` role	Indicates that the agent is the author of the lyrics or text of the work.
`has associated agent` qualified with `arranger` role	Indicates that the agent is credited with arranging or adapting the work.
`has associated agent` qualified with `producer` role	Indicates that the agent is responsible for the production of the recording and may hold neighbouring rights or sound recording copyright.
`performer` or `has associated agent` qualified with `performer`role	Indicates that the agent contributed to the recorded performance.

The composer creates the musical work. In the event-based CRM mapping:

E65 Creation → P14 carried out by → E21 Person
E21 Person → P94 has created → E28 Conceptual Object

A performer is an
E21 Person or an E74 Group, linked via P14 carried out by to the performance E7 Activity

A.5.2 Agent properties

Table A.12: Role and agent properties

Property	Explanation
has associated agent	Links an entity to an agent participating in its creation, performance, production, distribution, or documentation. The relationship must be qualified with subject has role to specify the agent's function. om:P19, skcmdb:P117, fuds:P453
(subject) `has role`	A qualifier for has associated agent that specifies the particular role performed by the agent. om:P20, skcmdb:P118, fuds:P463
`performer`	Records an explicit performance contribution when a simple performer property is needed instead of a role-qualified relationship. om:P152, skcmdb:P97, fu:P408
`record label`	The organisation credited as the imprint or label responsible for issuing a release. This refers to descriptive imprint information rather than rights ownership. |Equivalent to record label on Wikidata. For details see *Releases* earlier.
`date of birth`	Birth date of the agent, recorded when permitted by applicable data protection regulations.
`date of death`	Death date of the agent, recorded when permitted by applicable data protection regulations.
`ISNI`	The International Standard Name Identifier assigned to the agent. om:P92, skcmdb:P41, fu:461
`VIAF cluster ID`	The Virtual International Authority File identifier for the agent. om:P91, skcmdb:P40, fu:P60
`IPI Name Number`	The Interested Parties Information Name Number used for CISAC rights administration for composers, lyricists, and arrangers. om:P100, skcmdb:P42
`Spotify Artist ID`	A proprietary identifier used by Spotify for artists. om:P101, skcmdb:P43, fu:P61
MusicBrainz Artist ID	The identifier assigned to the agent in the MusicBrainz database. om:P103, skcmdb:P45, fu:P62
Discogs artist ID	The identifier assigned to the agent in the Discogs database. om:P105, skcmdb:P116, fu:P476

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

Stakeholder profile

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.

For the intake, we will also ask for a few lines of statement of interest in the Observatory Stakeholder Network and topic interests for data if they apply. We will make this information public, too.

We are also happy to create an intake interview and publish it on our website to allow the members of the Observatory Stakeholder Network to get familiar with each others ideas and interests.

> **!** Important
>
> We will provide in the final deliverable a link for the official intake to the stakeholder network. I would like to invite SOZA, Hudobné centrum, Aloaded, and HearDis! as first partners.

Open Music Data Exchange

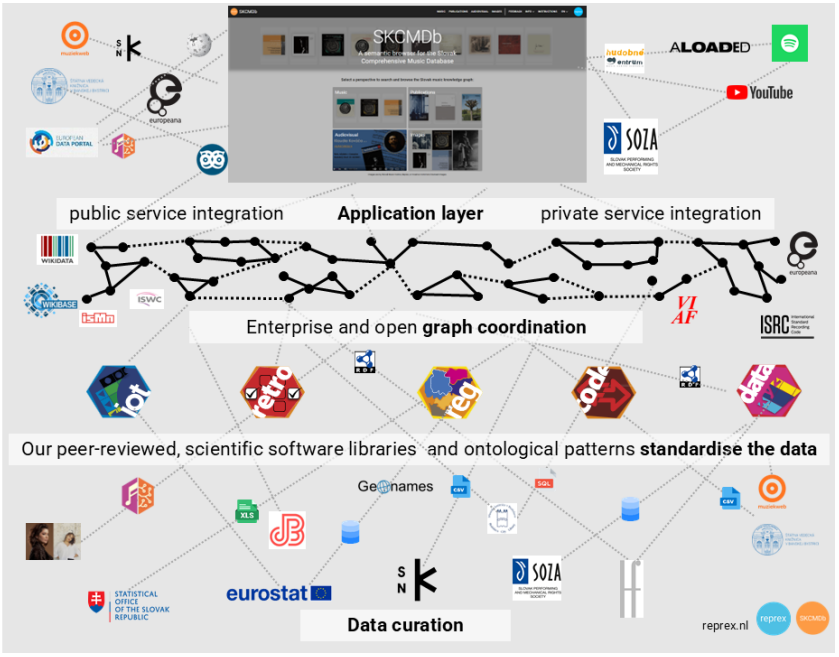
> **!** Important
>
> This advisory body will not be open for invitations. The representative pan-European stakeholders can nominate here their technical providers to consult the technical aspects of data exchange.

B SKCMDb: Slovak Comprehensive Music Database

The *Slovak Comprehensive Music Database* (SKCMDb) is a national initiative aimed at making Slovak music more accessible, discoverable, and usable across libraries, archives, streaming services, and rights management organisations. It connects scores, recordings, and metadata using open standards and collaborative governance. As a functional module of the Open Music Observatory, the SKCMDb also serves as a testbed for developing shared data services, addressing the conceptual models, workflows, and governance rules required to link diverse music stakeholders.

Dataspace for the European music sector

As a functional module of the Open Music Observatory, the SKCMDb also serves as a testbed for developing shared data services, addressing the conceptual models, workflows, and governance rules required to link diverse music stakeholders.



The SKCMDb is supported by a data-sharing space consisting of both shared and private databases.

The data-sharing space currently comprises the following initial components:

- **Slovak Metadata Database:** A database that facilitates connections between various Slovak stakeholders' systems.
- **SKCMDb (public):** A public database containing microdata on musical works, their recordings and scores, biographical and institutional information, statistical datasets, and a catalogue of publications.

- **`SKCMDb` HC-SOZA (private):** A database used exclusively for rights management and library management, governed by an agreement between the Slovak Music Centre and SOZA.
- **`SKCMDb` HF (private):** A technical dataset created to provide improvements, corrections, and enrichments for the Hudobný fond.

The *Slovak Metadata Database* serves as a support layer that is partly public and partly private. Its metadata definitions and descriptive metadata are exported into the `SKCMDb` databases as needed and permitted.

The Slovak Metadata Database

The Slovak Metadata Database is developed in alignment with the metadata framework of the Open Music Observatory.

- **Ontological and thesauri patterns:** Reuses standardized or widely adopted vocabularies.
- **Conceptualizations and definitions:** Includes concept definitions, thesauri, and other elements developed specifically for the `SKCMDb`.
- **Public permanent identifiers:** Uses identifiers that are public or can be made public.

The metadata layer is generally licensed under CC0, though in some cases other licenses are used (for example, CC-BY).

> **Note**
>
> **Examples:**
>
> - The definitions of *musical work*, *printed sheet music*, and the *is score of* relationship allow the description of connections between an abstract musical work—such as Bella's *Missa in C*—and its actual printed manifestations.
> - The VIAF identifier `2737220` identifies Ján Levoslav Bella's compositions across library systems.

Slovak Comprehensive Music Database (public)

The Slovak Comprehensive Music Database is a linked open database published by the Slovak Music Centre. It integrates elements from the Hudobné centrum's own databases along with data made public by SOZA, Hudobný fond, and other organizations. The database is distributed under various Creative Commons licenses that allow both commercial and non-profit use.

The primary aim of the SKCMDb is to reduce the cost of maintaining public and private services that enhance the circulation, availability, visibility, and legally licensed use of Slovak music. Our licensing policies are designed to enable the widest possible use of the data while protecting the investments required to maintain registers, standards, and data integrity.

Slovak Comprehensive Music Database (private)

The private components of the SKCMDb consist of databases where the data is not intended for public sharing but is used to enhance rights management, music information services, library operations, or other specialized applications. These databases are maintained under agreements between the participating parties.

Access to these private datasets serves specific, well-defined purposes and is governed by strict rules. Availability to third parties is determined solely at the discretion of the data owners and may vary depending on contractual or legal obligations.

B.0.1 Microdata

Microdata consists of information before it is aggregated into statistical datasets or formal publications. Metadata can also be considered microdata: while it is never aggregated, it plays a critical role in describing the provenance, semantics, and usability of aggregated data.

- **Collections:** Structured sets of similar items created through curatorial activities, where inclusion is based on discretionary selection to serve end users (e.g., a library's holdings or a curated playlist).
- **Registers:** Authoritative lists created through administrative processes with defined rules, aiming to capture all known items in a category (e.g., a national musical works register).

Collections typically rely on registers to identify works unambiguously and avoid duplication. Both are documented in structured datasets containing standard identifiers such as ISRC, ISWC, or ISMN codes. In the case of statistical data, microdata often refers to survey instruments and responses curated under defined methodological rules.

- **Metadata:** Relevant elements from the Slovak Metadata Database that support the use of collections or registers.

The SKCMDb's collections and register datasets are organized as a document database. This database stores structured data in RDF format describing musical works, sound recordings, printed and manuscript scores, as well as biographical information about music professionals and their organizations.

Each music-related object or agent (person, corporate body, or organization) is represented as a microdata dataset. These datasets share common definitions via conceptual models

and data structures, enabling automatic aggregation. All datasets are available with RDF annotation and can be exported in all standard RDF serializations.

Microdata is intended for institutional and professional use, not for the general public. It is annotated with standardized metadata suitable for applications such as music library cataloguing, distribution platforms, and rights management systems.

> **Note**
>
> **Example**
> The musical work *String Quartet in B-flat Major* (composed by Ján Levoslav Bella) is described in a dataset that includes two publicly available printed scores and a publicly available sound recording.
>
> - **Graphical view:** Navigate and contribute to detailed entries on musical works, sound recordings, and related assets. Explore on Wikibase
> - **Semantic view:** Export structured data in XML, JSON-LD, Turtle, or N-Triples formats for reuse in research or digital projects. Example Turtle file or download XML
>
> Recording of microdata follows defined rules. As a general principle, living natural persons may opt out of inclusion in the database.

B.0.2 Statistical Data & Data Catalogue

Our statistical data and catalogue consist of datasets aggregated using statistical methodologies. These comply with the SDMX standard and the W3C Data Cube vocabulary, making them compatible with spreadsheet software, statistical packages, and data science workflows in R, Python, or similar environments.

Datasets are offered in multiple formats. In addition to RDF serializations, we provide standard CSV files and, when required, Excel or SPSS formats. We also publish data papers and related documentation that describe dataset usability and highlight key insights.

B.0.3 Publications & Catalogue

The `SKCMDb`'s most important publications are musical works, made available to end users as sound or video recordings and printed sheet music. These may be distributed in different sales formats, such as physical albums or books.

Microdata and statistical datasets are treated as publications and are listed both in the general catalogue and a machine-readable data catalogue.

The `SKCMDb` also includes methodological and musicological publications, as well as data papers explaining the use of datasets. The catalogue is designed for interoperability with libraries, archives, museums, and similar institutions.

In some cases, the SKCMDb may host the full publication, with the Open Music Observatory acting as publisher. In most cases, however, it provides catalogue entries with clear access points, such as webshops, public library lending systems, or repository links where legal copies of the documents can be obtained.

C LīvMDb: Livonian Music Database {#sec-annex-livmdb .unnumbered, .appendix}

The *Livonian Music Database* (`LīvMDb`) is a proof-of concept for working with music that has low documentation depth, weak institutions. The music of the Livonian people is scattered, and as native speakers of this small ethnic group died out, their heritage was dispersed, and largely not placed on modern digital platforms. The `LīvMDb` as a functional module of the Open Music Observatory, serves as a testbed for working with very low documentation, decolonisation, and other issues related to regional cultures, ethnic minorities. It offers insights into subsidiarity.

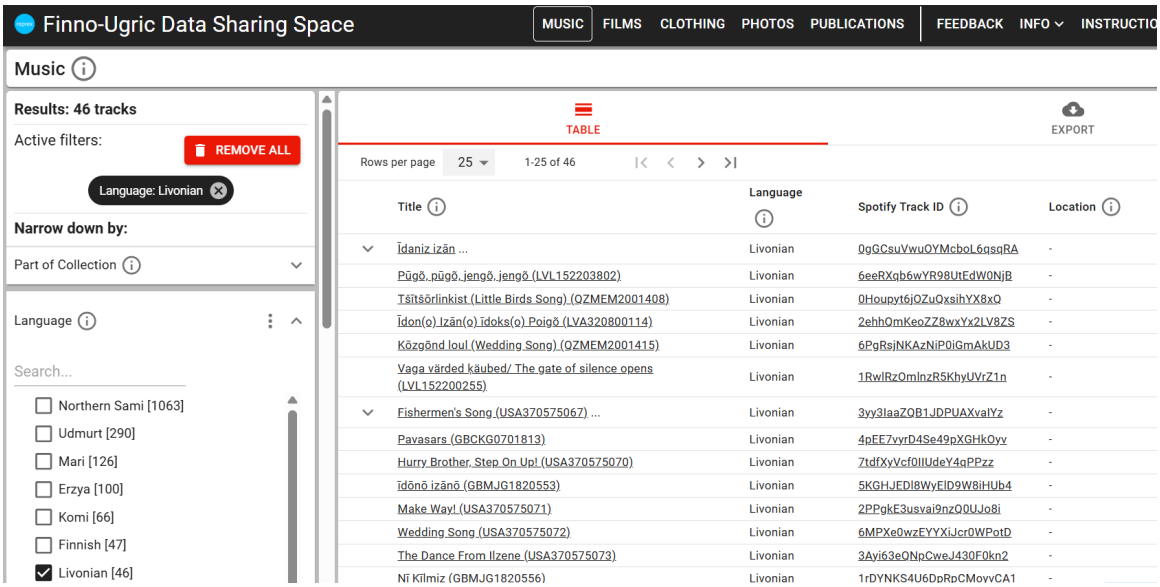


Figure C.1: The LīvMDb is federated with the Finno-Ugric Data Sharing Space and the Open Music Observatory. The first places Livonian music into a wider Finno-Ugric cultural context, the latter into a music context.

The `LīvMDb` is supported by a data-sharing space consisting of both shared databases.

The data-sharing space currently comprises the following initial components:

- **Finno-Ugric Metadata Database:** A database that facilitates connections between various Finno-Ugric and Baltic stakeholders' systems.

- **`LīvMDb` (public):** A public database containing microdata on musical works, their recordings and scores, biographical and institutional information, statistical datasets, and a catalogue of publications.
- **`LīvMDb` (private):** A technical dataset created to provide improvements, corrections, and enrichments.

The *Livonian Metadata Database* serves as a support layer that is partly public and partly private. Its metadata definitions and descriptive metadata are exported into the `LīvMDb` databases as needed and permitted.

The Livonian Metadata Database

The Livonian Metadata Database is developed in alignment with the metadata framework of the Open Music Observatory.

- **Ontological and thesauri patterns:** Reuses standardized or widely adopted vocabularies.
- **Conceptualizations and definitions:** Includes concept definitions, thesauri, and other elements developed specifically for the `LīvMDb`.
- **Public permanent identifiers:** Uses identifiers that are public or can be made public.

The metadata layer is generally licensed under CC0, though in some cases other licenses are used (for example, CC-BY).

> **Note**
>
> **Examples:**
>
> - The definitions of *musical work*, *music recording* and *is recording of* relationship allow the description of connections between an abstract musical work and its recording(s).
> - The *ISRC code* identifies recordings across all streaming platforms.

Livonian Music Database (public)

The Livonian Music Database is a linked open database published by the Reprex on behalf of the Open Music Observatory. The database is distributed under various Creative Commons licenses that allow both commercial and non-profit use.

The primary aim of the `LīvMDb` is to provide a use case for very low documentation music ecosystems with very limited resources and challenging data curatorial scenarios.

Livonian Music Database (private)

The private components of the LīvMDb are a staging area for data that has unclear provenance or legal status. Unlike in the case of LīvMDb, we hold minimal business confidential data (related to the royalty accounts of music that we published), but some data may have, for example, unclear GDPR status.

C.0.1 Microdata

Microdata consists of information before it is aggregated into statistical datasets or formal publications. Metadata can also be considered microdata: while it is never aggregated, it plays a critical role in describing the provenance, semantics, and usability of aggregated data.

- **Collections:** Structured sets of similar items created through curatorial activities, where inclusion is based on discretionary selection to serve end users (e.g., a library's holdings or a curated playlist). Our work is centerred around the work of *Hõimulõimed*, a Finno-Ugric NGO, which curated many Finno-Ugric language collections, including the collection of Livonian-language songs available on Spotify.
- **Registers:** Authoritative lists created through administrative processes with defined rules, aiming to capture all known items in a category. Our work buids on the register of Livonian placenames, because they offer the most straighforward curatorial help to find new Livonian (folk) music[1].
- **Metadata:** Relevant elements from the Livonian Metadata Database that support the use of collections or registers.

The LīvMDb's collections and register datasets are organized as a document database. This database stores structured data in RDF format describing musical works, sound recordings, printed and manuscript scores, as well as biographical information about music professionals and their organizations.

Each music-related object or agent (person, corporate body, or organization) is represented as a microdata dataset. These datasets share common definitions via conceptual models and data structures, enabling automatic aggregation. All datasets are available with RDF annotation and can be exported in all standard RDF serializations.

Microdata is intended for institutional and professional use, not for the general public. It is annotated with standardized metadata suitable for applications such as music library cataloguing, distribution platforms, and rights management systems.

[1] *Livonian place names: documentation, problems, and opportunities* (Ernštreits 2020) and our gazetteer dataset: (Antal, Pigozne, and Mester 2025).

Note

Example
The village of *Mazirbe* (Livonian: Irē, German: Klein-Irben, Russian:) is the central location of the Livonian culture, which hosts the Livonian Community House. Folk songs collected and recordings made in Mazirbe may have a provenance of Irē, Mazirbe, Klein-Irben (or its Finnish and Estonian versions), or various Cyrillic transliterations. Mainitaing a clear metadata dataset on the geographical sources of Livonian music is necessary.

- **Graphical view:** Navigate and contribute to detailed entries on musical works, sound recordings, and related assets. Explore on Wikibase
- **Semantic view:** Export structured data in XML, JSON-LD, Turtle, or N-Triples formats for reuse in research or digital projects. Example Turtle file or download XML

We made available the dataset in standard RDFXML, JSON-LD, and TTL serialisations accompanies by a data paper explaining its use (Antal et al. 2025; Antal, Pigozne, and Mester 2025)

C.0.2 Statistical Data & Data Catalogue

Our statistical data and catalogue consist of datasets aggregated using statistical methodologies. These comply with the SDMX standard and the W3C Data Cube vocabulary, making them compatible with spreadsheet software, statistical packages, and data science workflows in R, Python, or similar environments.

Datasets are offered in multiple formats. In addition to RDF serializations, we provide standard CSV files and, when required, Excel or SPSS formats. We also publish data papers and related documentation that describe dataset usability and highlight key insights.

C.0.3 Publications & Catalogue

The `LīvMDb`'s most important publications are sound recordings, which are made available as a archival recordings (not ready to be communicated to the public on commercial platforms) and publicly available sound recordings. We also provide access to some printed and and hand-written scores.

Microdata and statistical datasets are treated as publications and are listed both in the general catalogue and a machine-readable data catalogue.

The `LīvMDb` also includes methodological and musicological publications, as well as data papers explaining the use of datasets. The catalogue is designed for interoperability with libraries, archives, museums, and similar institutions.