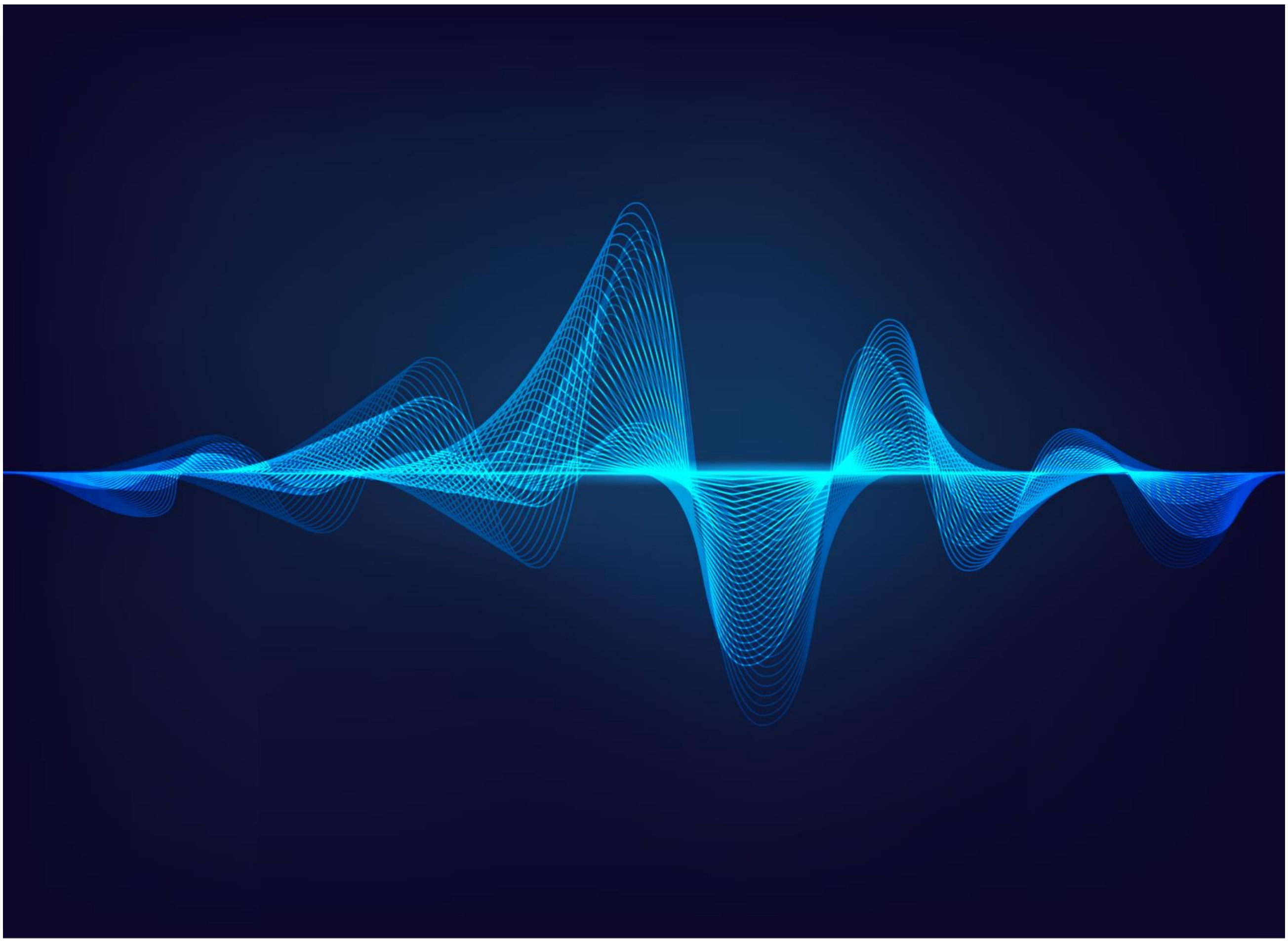

D4.1 – Open-Source Music Data Software Ecosystem

OpenMusE

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

Funded by the European Union

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

Project Factsheet

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

Deliverable Factsheet

Number:	**D4.1**
Title:	**Open-Source Music Data Software Ecosystem**
Lead beneficiary:	UTU
Work package:	WP4
Task:	T4.1
Dissemination level:	Public
Submission date (v1):	28.02.2024
Resubmission (v1.1):	17.07.2024
Main author(s):	Leo Lahti, Pyry Kantanen, Jeba Akewak (UTU)
Contributor(s):	SINUS, UTU, UVA, SSSA, ALOADED, HEARDIS
Quality check:	Richard Wages, SINUS James Edwards, SINUS

Document history:

Revision	Date	Main modification	Author
0.1	25/07/2023	Headlines & Index fixed	Leo Lahti
0.5	21/2/2024	Report draft	Pyry Kantanen
0.8	23/2/2024	Final draft	Leo Lahti
0.9	28/2/2024	Final version	Leo Lahti
1.0	28/2/2024	Submitted to EC	Richard Wages
1.1	15/07/2024	V1.1 prepared in response to Project Officer and Reviewer feedback	Leo Lahti

Disclaimer of Warranties

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

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

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

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

(a) makes any warranty or representation whatsoever, express or implied,

  (i). with respect to the use of any information, apparatus, method, process, or similar item disclosed in this document, including merchantability and fitness for a particular purpose, or

  (ii). that such use does not infringe on or interfere with privately owned rights, including any party's intellectual property, or

  (iii). that this document is suitable to any particular user's circumstance; or

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

Glossary

API Application programming interfaces

CRAN Comprehensive R Archive Network, central software repository for R and R packages and their documentation

DMP Data Management Plan

DOI Digital Object Identifier, "a persistent identifier or handle used to uniquely identify various objects"[1]. Fits within the URI system

Documentation (Software) Written text or illustration that accompanies computer software. Explains in general terms how the software operates and how the software can be used.

Eurostat Eurostat is the Statistical Office of the European Union. Eurostat does not directly collect data itself, apart from a small number of exceptions. Data collection is done in EU countries by national statistical authorities in compliance with common EU statistical regulations and standards.

EU-SILC Statistics on Income and Living Conditions. A survey-based, harmonised statistics on income and living conditions in the EU.

FAIR Findable, Accessible, Interoperable, and Re-usable data

Free and open-source software (FOSS) Software that is both open-source but also freely licensed, meaning that everyone can modify and redistribute the source code of the software without significant restrictions.

Git "A distributed version control system that tracks changes in any set of computer files, usually used to coordinate work among programmers who are collaboratively developing source code during software development."[2]

GitHub A platform commonly used to host open-source software development projects using Git distributed version control.

Issue (GitHub) A question, bug report, suggestion, or other type of communication that users of a specific software or software component can open in a GitHub repository or other similar software repository

JSON-LD JavaScript Object Notation for Linked Data, a method of encoding linked data using JSON.

License (software) Software license is a document that states the rights of the developer and the user for a given software.

Metadata A statement about a potentially informative object, usually on how an informative objects such as a dataset or file can be found, accessed, used.

Open-source Open-source software is software which source code is published and made available to the public, enabling anyone to copy, modify and redistribute the source code, design documents, or content of the product.

R An open-source statistical environment and an accompanying, high-level computer language that mainly aims to support programmatic and reproducible data management and statistical work.

[1] Wikipedia contributors. (2024, February 20). Digital object identifier. In *Wikipedia, The Free Encyclopedia*. Retrieved February 23, 2024, from https://en.wikipedia.org/w/index.php?title=Digital_object_identifier&oldid=1209088493

[2] Wikipedia contributors. (2024, January 1). Git. In *Wikipedia, The Free Encyclopedia*. Retrieved February 23, 2024, from https://en.wikipedia.org/w/index.php?title=Git&oldid=1193006801

R package additional software component to R that supplements the built-in software components by adding additional features and functionalities.

Repository (software) centrally located storage where software project's files, documentation, and other resources can be stored. A repository also includes project history, making it possible to view and access older versions of software source code. Repositories can also have platforms for opening issues, making pull requests or other code contributions, opening discussions, and monitoring contributor activities.

RDF Resource Description Framework, a standard for data interchange in web.

SDMX Statistical Data and Metadata Exchange, 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 member countries. Sponsored by Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat (Statistical Office of the European Union), International Monetary Fund (IMF), Organisation for Economic Cooperation and Development (OECD), United Nations Statistical Division (UNSD), and World Bank.

Software-as-a-Service, SaaS In generic terms refers to licensing and delivery model of a centrally hosted software solution, but here refers to centrally hosted web-based software.

Semantic versioning, Semver software versioning system often consisting of 3 integers separated by dots. The first integer denotes a major version of the software, the second integer denotes a minor version of the software, and the third integer denotes a patch to the software. When software version changes from one major release to another, the changes are usually more drastic than when the software releases minor revisions or patches and bug fixes.

Shiny app web application framework for R, designed to turn analyses into interactive web applications.

Statistics quantitative and qualitative, aggregated, and representative information characterising a collective phenomenon in a considered population.

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.

Tidy data data structured in a way where "each variable is a column, each observation is a row, and each type of observational unit is a table" (Wickham 2014).

Tidy workflow a reproducible data science workflow supporting tidy data principles.

Turtle Terse RDF Triple Language is a syntax and file format for expressing data in the Resource Description Framework (RDF) data model.

URI Uniform Resource Identifier, "a unique sequence of characters that identifies an abstract or physical resource"[3]

URL Uniform Resource Locator, a specific type of URI that is a reference or address used to access a resource on the internet.

[3] Wikipedia contributors. (2024, February 23). Uniform Resource Identifier. In *Wikipedia, The Free Encyclopedia*. Retrieved February 23, 2024, from https://en.wikipedia.org/w/index.php?title=Uniform_Resource_Identifier&oldid=1209684355

Table of Contents

Executive summary

WP4 seeks to develop an open-source, software-as-a-service (SaaS) collection of tools for music data, improving and integrating previously initiated software from the partners and augmenting this with new software components. This supports the OpenMusE goal to pioneer new best-practice methods and tools for data collection from multiple sources and integrate these into an open-source software ecosystem that non-specialist stakeholders can use. The data needs identified together with WP1-WP3 guided software modifications and extensions, and the creation of new software components as outlined in the project proposal.

Within WP4, T4.1 focuses on software for the collection and management of statistical data. It improves interoperability of the software with the latest database APIs and recommended data standards outlined in the DMP, securing software sustainability, and including a better tracking for data citation capabilities that is essential for data provenance. The software provides tools to check and amend metadata based on existing, publicly available datasets accessed through the prior packages and a general strategy to deal with emerging needs for new data. Novel cloud-based applications with web interfaces (R Shiny Apps) and further interface improvement have been implemented for usability improvement for non-scientific users. The work is delivered through open-source repositories with usability documentation; this report summarises the main elements of the deliverable.

1 Introduction

A key *objective* of the OpenMusE project is to provide an open-source toolkit for music data. In the context of this project, *music data* refers to a variety of data types relevant to music consumption, including more traditional statistical, economical, administrative, survey data as well as music streaming data from both public and private (e.g. partner-provided) sources. WP4 focuses on statistical and survey data collection and management tools. In parallel, tools for collecting and managing music industry accounting system data, music streaming data, and music metadata will be developed in WP1, WP2, and WP5. The WP teams will work with common standards to ensure interoperability across the toolkit. The resulting toolkit will represent a novel combination of tools with added value and a general strategy for further extension.

D4.1 is the result of the work done in Task 4.1. It builds on the data needs identified with WP1-3 (in particular, Tasks T1.1, T2.1, T3.1). Deliverable D4.1 focuses on the collection and management of statistical data available on an EU level via Eurostat. Specifically, it updates the *eurostat* R package (https://cran.r-project.org/web/packages/eurostat/index.html) to allow interoperability with new database APIs and provide new data citation functionality. T4.1 / D4.1 will be followed and complemented by tasks and deliverables focusing on the collection and management of primary survey data.

2 Background work

2.1 Data gaps

The work on D4.1 builds on the data needs identified with WP1-3 and in particular, Tasks T1.1, T2.1, and T3.1. The data needs were mapped based on the deliverables that were available from these tasks, complemented by additional in-person interviews with WP1-3 during fall 2023. DMP Annex 2 provides an additional listing of relevant data sources. These are complemented by *user stories, or* layman descriptions of common use cases, supporting the development of user-friendly software solutions; REPREX maintains a publicly available collection of user stories[4] that is used to support the development work in WP4 (as well as in parallel development tasks in WP1, 2, 3, and 5). These specific user scenarios must be reconciled with the need to provide more widely applicable methods that can

[4] The main Github repository for the user stories
https://github.com/dataobservatory-eu/open-music-europe-user-stories/tree/main/stories
DMO Personas document includes additional user stories:
https://docs.google.com/document/d/185IeCTRnjVv4lY3kYJQKJ1kD93V3ubfFwtNH3wWv_ak

efficiently serve varying use cases and data types relevant to WP1-3. It has been therefore essential to focus on generic methods that support data provenance, metadata augmentation, and compliance with established data standards. The added value of D4.1 is specifically in assisting the integration heterogeneous data sources that have varying compliance with existing standards. Adapting this approach to specific data sources and use cases will continue in the remaining WP4 tasks, as well as in tasks in WP1, 2, 3, and 5 that implement the indicators, methods, and *software tools* designed by OpenMusE.

2.2 Updates to the original work plan

Updates to planned work have been inevitable as the work progresses. The main changes include shifting focus from package dependencies to interoperability through workflow design and prioritizing the development of new workflows and applications over previous background components. Major updates to T4.1 included the following.

1. Emphasis was shifted from interoperable software to implementing modular workflows as some of the technical overheads on software dependencies can be overcome with this development strategy.
2. Eurostat deprecated its old API in October 2023 and substantial unexpected work was needed to upgrade the *eurostat* package (Lahti et al. 2017 & 2024) to support the new SDMX API and maintain functionality.
3. The latest terms of using data from Spotify API with the *spotifyr* package are too limiting for the open data science framework envisioned for this consortium. T4.1 assessed the possibilities for replacing this with similar alternative data sources (e.g. Deezer, MusicBrainz), and may support as needed the streaming data collection and management activities in WP1.
4. T4.1 decided to reduce software dependencies by shifting the emphasis from creating package dependencies to implementing integrative data science (Quarto) workflows and an interactive cloud-based (Shiny) application (e.g. Fig. 1). This demonstrates a general strategy for integrating previous and newly created open components into a new combination that has unique added value facilitating music data access.

2.3 Elements for continuation

The main elements of the follow-up development will comprise (i) improving the existing software (critical updates with API changes, bug fixes, enhanced usability etc.); (ii) customizing the generic solutions provided by the toolkit developed in T4.1 to support the use of specific data sources and data sets in WP1, 2, 3, and 5; (iii) expanding the toolkit to ingest an increasing number of data sources; and (iv) creating suitable instructional material (vignettes, walkthroughs, etc.) on how to use the tools presented in D4.1 to access music data in specific.

3 Software

OpenMusE seeks to develop an open-source, software-as-a-service (SaaS) toolkit for music data collection. In the context of this project, *music data* broadly refers to a variety of data types relevant to music, including more traditional statistical, economical, administrative, survey data as well as music streaming data from both public and private (e.g. partner-provided) sources.

D4.1 focuses on the collection and management of EU-level statistical data. It was designed to *improve and integrate* previously initiated software brought in by the partners and augment them with other open-source components and newly created components. T4.1 evaluated alternative strategies for enhancing the interoperability and joint use of the software, the compatibility of input and output data with recommended formats (outlined in DMP), and the need to create user-friendly interactive applications around these methods (Fig. 1).

D4.1 is delivered as a collection of open-source components and application built around them. In addition to offering a ready-to-use tool, it establishes a development strategy that can be extended to a number of data sources and combinations.

3.1 Components

Collectively, the software tools developed in T4.1 provide enhanced methods and data science strategies to access, process, and analyse statistical data on music. The selected strategy of combining the software elements at the level of workflows is exemplified by the workflow documented at the eurostat interactive Shiny app, which jointly utilises a number of open-source components.

The interactive application combines openly available software components into a coherent workflow supporting a variety of data formats (Fig. 1). This approach provides a data science strategy that could be extended or replicated to include a growing number of data sources and their combinations. Links to additional open-source components that D4.1 could take advantage of are listed in the *Open Music Software Ecosystem* vision paper (Antal 2024c). The software packages from REPREX are openly licensed, allowing their mixing and reuse with other software.

D4.1 represent a unique combination of open-source software from UTU as well as from external developers as described in the original T4.1 plan, and a general data science strategy exemplified by the work based on the *eurostat* package. In the following, we summarise the main components of the software toolkit and additional, closely linked software whose open-source licensing terms allow flexible use and development in support of D4.1.

Shiny application (UTU): interactive access to Eurostat open data

- URL: https://pitkant.shinyapps.io/stats_shiny/
- Updates: new software component that demonstrates the data science strategy in D4.1

eurostat (UTU): programmatic access to Eurostat open data

- URL: https://github.com/rOpenGov/eurostat (also available in CRAN)
- Updates: API upgrade; data format support; see the GitHub commit log

RShiny: web framework for building web applications using R

- URL: https://www.rstudio.com/products/shiny/
- External package: not developed in T4.1 but providing essential complementary methods and used to create the interactive application to support data access and use.

tidyverse: collection of R packages designed for data science

- URL: https://www.tidyverse.org/
- External package: not developed in T4.1 but providing essential complementary methods and used to create the interactive application to support data access and use.

3.2 The *eurostat* Package

Let us next demonstrate how the toolkit can be put in practice to jointly utilise different software components in specific data retrieval use cases that T4.2 and T5.1 will focus on. The *eurostat* package addresses identified needs for specific *statistical* and *survey data sources.* For instance, it provides access to EU Statistics on Income and Living Conditions (**EU-SILC**[5]) which is managed by Eurostat and can be accessed easily through the eurostat package. EU-SILC serves as a critical source of data, primarily concentrating on topics such as income, poverty, social exclusion and living conditions, offering a detailed insight into the material well-being of citizens across the European union region. The *eurostat* package can be used to download this and other data sets from the Eurostat data portal in a tidy data format but the overall data processing workflow benefits from combining the data retrieval tool with other methods that provide enhanced capabilities for data provenance, metadata augmentation, and downstream data processing (Fig. 1).

The *eurostat* package provides access to the Eurostat open data warehouse. During M7-M13 of the project, a major new release was developed and released. Large portion of the updates related to changes in the Eurostat API; deprecation of the old bulk download facilities and the old JSON web service and moving to use the new SDMX API and API Statistics. Thus, implementing the changes kept the software functional with critical open API updates. In addition to the necessary changes needed to keep the package functional, the new major release version 4.0.0 presented the opportunity to address long-standing feature requests, bug reports, quality-of-life improvements, and other issues. The shift from minor and patch versions where backwards compatibility had to be maintained to releasing a major version where backwards compatibility could be slightly compromised provided an opportunity to make larger changes than would usually be possible[6]. The change from old bulk download facility and JSON web service were handled to preserve backwards compatibility as much as possible. The format of the data output for data retrieval functions are almost identical to version 3.8.3, with the largest differences visible to the end user being related to using a new naming convention for the time column and including a new column indicating the frequency at which the data was collected.

3.2.1 Summary of the upgrades

Technical details of the software development steps are available in the respective GitHub repositories of each package. Here, we summarise changes related to the *eurostat* package. This exemplifies the development work that is needed to maintain software sustainable, enhance interoperability with data standards and other software components, improve user experience, and support the development of interactive applications.

Most significant changes in the *eurostat* package were targeted for improved user experience in terms of reliability, provenance, accessibility, and speed. We improved dataset caching functionalities, reducing unnecessary traffic between Eurostat API and end user by using cached datasets if the retrieved dataset is a subset of a previously retrieved dataset, which leads to essential speed-up from ordinary user perspective. We added folder hierarchy handling to functions related to Eurostat Table of Contents, making it possible to download all datasets that belong to a certain folder with a new function, also a change that will substantially contribute to an improved user experience. We made improved data references and provenance by time, version and output format information that improve compatibility with external tools such as LaTeX renderers. We also include citation to the eurostat software package in the package CITATION in addition to the previously included software paper citation to better conform with FORCE11 Software Citation Principles (Smith, Katz & Niemeyer

[5] https://ec.europa.eu/eurostat/web/main/home
[6] See Semantic Versioning 2.0.0 standard: https://semver.org

2016). We added support for German and French, in addition to the previously supported English, after which it is possible search for datasets in Eurostat Table of Contents in multiple languages, label variable names and categorical variables in datasets in multiple languages and print dataset citations that include dataset names in the language of choice. We replaced dependencies on the *httr* (Wickham 2023) package with *httr2* package (Wickham 2023), which is more modern reimplementation and provides futureproofing for the *eurostat* package's essential functionalities, added the option to use the *data.table* package (Barrett et al. 2024), which is adds speed and reduces RAM use when dealing with large datasets with hundreds of millions of rows, and aimed to make minimal installations without the *sf* package (Pebesma 2018) possible by rewriting map drawing functionalities in the *eurostat* package and off-loading functions to a separate package, the *giscoR* package (Hernangómez 2024). We made it possible to search Eurostat Table of Contents from the code field in addition to the default title field. We also enhanced automated documentation to reduce potential errors and improve quality (see Wickham & Bryan 2023). We also added a new interactive data retrieval function that walks the user through the following steps: finding dataset, selecting download parameters, downloading a dataset, printing a data citation for the dataset, and, following the recommendations found in the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group 2014), printing a fixity sum for each dataset.

3.2.2 Eurostat statistics browser (interactive Shiny application)

The Eurostat statistics browser Shiny application uses the *eurostat* package to retrieve data from the Eurostat API. The data retrieval parameters and be set and the retrieved dataset can be viewed in an interactive Shiny application (see Chang W et al. 2023). The process is summarised in Fig. 1. The datasets can be enriched by using the *dataset* package (Antal 2024a), which is not part of D4.1 but its open licensing conditions allow reuse with other software. The user can add metadata to the downloaded dataset by converting the downloaded dataset to a special dataset class with the contents of the metadata fields coming directly from the Eurostat SDMX API. Additionally, datasets with dataset class can be converted to DataCite and DublinCore formats, with fields following the DataCite and DublinCore conventions (see Antal 2023a). Additionally, datasets with these annotations can be converted to other linked data formats, such as Turtle (the *tuRtle* package from REPREX, Antal 2024b, could be developed further to provide enhanced support for such tasks), RDF and JSON-LD, with the help of *redland* (Jones et al 2023) and *rdflib* (Boettiger 2018) packages. Including metadata with the downloaded datasets makes it easier to implement FAIR (Findability, Accessibility, Interoperability, Reusability) principles in downstream analyses and data dissemination.

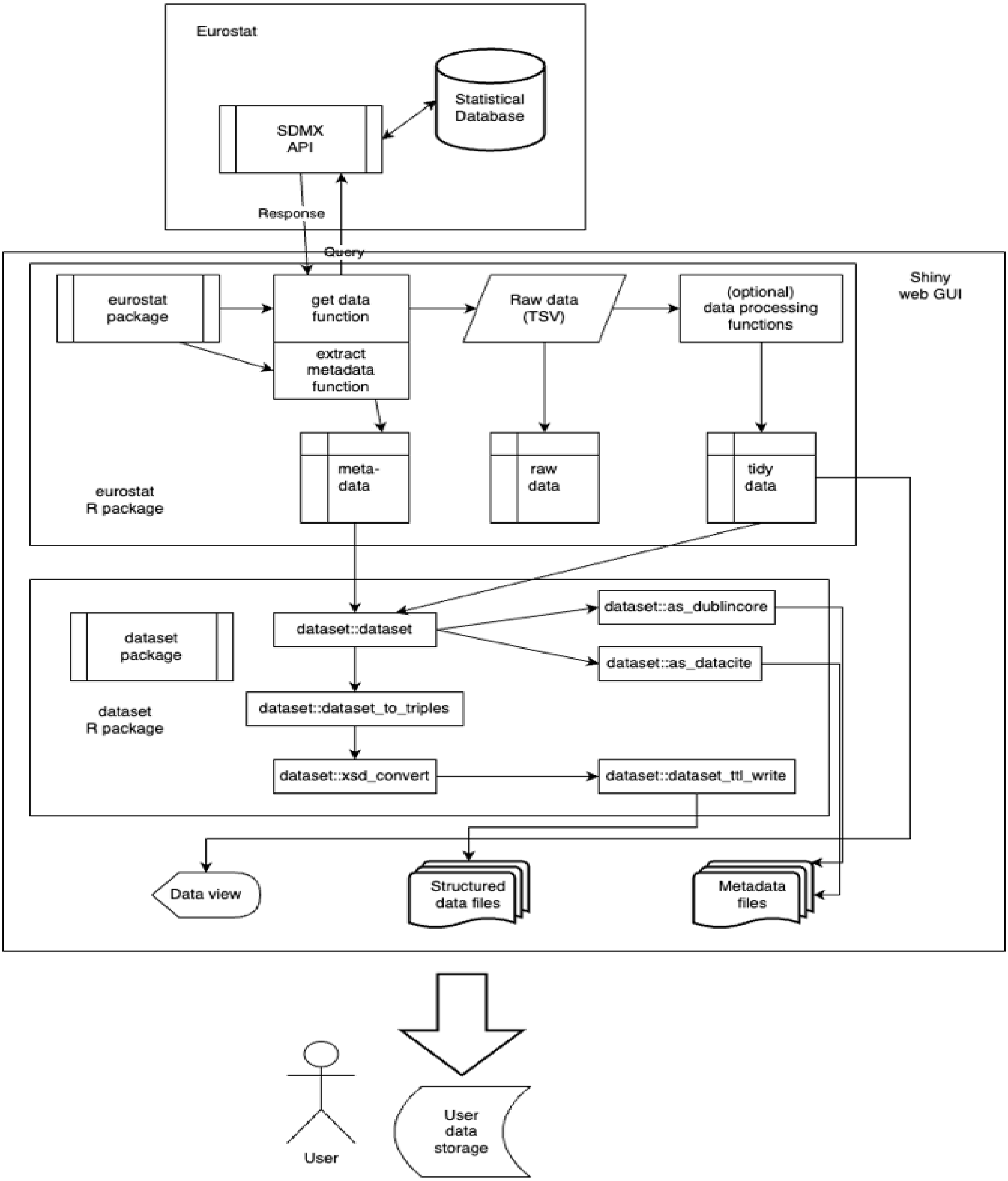

Figure 1: Overview of data retrieval and metadata processing.

The source code of the Eurostat statistics browser Shiny app is openly licensed and available in GitHub with instructions on how to run it locally[7]. For user convenience, the app is also available for testing in

[7] https://github.com/pitkant/stats_shiny

shinyapps.io[8], but it can be deployed in any environment. The figure describes the architecture of the interactive application.

3.2.3 Potential extensions

T4.1 evaluated the need for new software components to complement improvements on the previously developed set of packages. Assessment of software dependencies led to the conclusion that integrating functionalities through reproducible workflows and applications is the preferred strategy that allows scaling up interoperability between independently developed open-source components; this minimises dependencies between individual packages keeps codebases easier to maintain and limits adverse consequences when individual components are broken. In practice this means preference to software that focus on a handful of essential features and produce predictable outputs that can be handled by other software. Interoperability can be realised by writing additional software that orchestrates the input and output of different components and hides details that would be unnecessary to most end users. This takes the form of a reproducible *Quarto document* (tutorials available at the software homepages compile in Quarto) and an interactive *Shiny application* released through a cloud service. WP4 chose this strategy as the basis for development work. Thus, we developed OpenMusE interactive cloud-based online service (*Shiny app;* by UTU, based on openly licensed software components) to improve usability for non-scientific users.

The toolkit could benefit from additional components to strengthen the support for alternative data formats and data sources. The *Open Music Europe Software Ecosystem* vision paper (Antal 2024c) provides suggestions towards this direction; for instance, in principle, the *dataset* and *tuRtle* packages from REPREX could be used as dependencies for music data collection software. We note that these components can be flexibly integrated also in the follow-up work under open-source licensing conditions.

[8] https://pitkant.shinyapps.io/stats_shiny/

Conclusion

OpenMusE seeks to develop a software toolkit for the collection and integration of music data from various sources. The work is delivered through open-source repositories as summarised in this report. The data needs identified with WP1-3 regarding music data include statistical, survey, and streaming data sources. D4.1 focuses on EU-level statistical data, providing tools that represent a unique combination of previous and newly created open-source software and a demonstrated strategy for creating integrative workflows and interactive applications that make these methods accessible to a broader audience including both data scientists and non-technical users.

T4.1 has evaluated alternative strategies for enhancing the interoperability and joint utilisation of the software components, compatibility of their input and output data with the recommended formats outlined in the DMP, and user-friendly interactive applications supporting the use of these methods. Changes were implemented to improve the interoperability of the software with the latest database APIs and recommended data standards outlined in the DMP, securing software sustainability, and including a better tracking for data citation capabilities that is essential for data provenance; the online resources demonstrate how these tools can generically be used to check and amend metadata based on existing, publicly available datasets. Assigning correct and sufficient metadata schemas, properties and namespaces to each dataset remains the responsibility of expert data curators, and the adopted data science strategy can be replicated to provide customised solutions for an increasing number of data sources and combinations.

Future development will be facilitated by the open-source approach that has been chosen as the basis for OpenMusE development. In addition to maintaining and improving the existing software, taking advantage of available open-source solutions, and customizing the generic solutions to work more closely on specific data sources identified in WP1, 2, 3, and 5, the toolkit can be expanded to ingest an increasing number of new types of relevant statistical, survey, and streaming data. Technical variability in the data types and formats poses continuing challenges for the prioritisation of the multitude of the dozens of heterogeneous sources of relevant data to support variable use cases. Interface improvements and the creation of new interactive applications following our suggested strategy will create added value on the openly available software components and can help to improve the usability of these methods among both practicing data scientists as well as non-technical users.

References

Antal D (2024a). dataset: Create Data Frames that are Easier to Exchange and Reuse. R package version 0.3.0, https://CRAN.R-project.org/package=dataset

Antal D (2024b). tuRtle: Parse and Export R Data with the Turtle Syntax for the Resource Description Framework. R package version 0.1.0, https://dataobservatory-eu.github.io/tuRtle/

Antal D (2024c). Open Music Europe Software Ecosystem. DOI: 10.5281/zenodo.10578359

Antal D (2023a). Motivation: Make Tidy Datasets Easier to Release Exchange and Reuse. Referenced 2024-02-22, URL: https://dataset.dataobservatory.eu/articles/motivation.html

Antal D (2023b). "Reproducible Input-Output Economics Analysis, Economic and Environmental Impact Assessment with Empirical Data." DOI: 10.5281/zenodo.5887037, https://iotables.dataobservatory.eu/

Antal D (2021a). retroharmonize: Ex Post Survey Data Harmonization. R package version 0.2.0, https://CRAN.R-project.org/package=retroharmonize

Antal D (2021b). regions: Processing Regional Statistics. R package version 0.1.8, https://CRAN.R-project.org/package=regions

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T (2024). data.table: Extension of data.frame. R package version 1.15.0, https://CRAN.R-project.org/package=data.table

Boettiger C (2018). rdflib: A high level wrapper around the redland package for Common rdf applications (Version 0.1.0). Zenodo. https://doi.org/10.5281/zenodo.1098478

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023). shiny: Web Application Framework for R. R package version 1.7.5.1, https://CRAN.R-project.org/package=shiny

Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11, https://doi.org/10.25490/a97f-egyk

Hernangómez D (2024). giscoR: Download Map Data from GISCO API - Eurostat. DOI: 10.5281/zenodo.4317946, https://ropengov.github.io/giscoR/

Jones M, Slaughter P, Ooms J, Boettiger C, Chamberlain S (2023). redland: RDF Library Bindings in R. DOI: 10.5063/F1VM496B, R package version 1.0.17-17, https://github.com/ropensci/redland-bindings/tree/master/R/redland

Lahti L, Huovari J, Kainu M, Biecek P, Hernangomez D, Antal D, Kantanen P (2024). eurostat: Tools for Eurostat Open Data [Computer software]. R package version 4.0.0. https://CRAN.R-project.org/package=eurostat

Lahti L, Huovari J, Kainu M, and Biecek P (2017). Retrieval and analysis of Eurostat open data with the eurostat package. The R Journal 9(1), pp. 385-392. doi: 10.32614/RJ-2017-019

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group.
(2016) Software Citation Principles. PeerJ Computer Science 2:e86.
DOI: 10.7717/peerj-cs.86

Thompson C, Antal D, Parry J, Phipps D, Wolff T (2022). spotifyr: R Wrapper for the 'Spotify' Web API. R package version 2.2.4, https://CRAN.R-project.org/package=spotifyr

Wickham, H (2014). Tidy Data. Journal of Statistical Software, 59:10. DOI: 10.18637/jss.v059.i10

Wickham H (2023). httr: Tools for Working with URLs and HTTP. R package version 1.4.7, https://CRAN.R-project.org/package=httr

Wickham H (2023). httr2: Perform HTTP Requests and Process the Responses. R package version 1.0.0, https://CRAN.R-project.org/package=httr2

Wickham H & Bryan J (2023). R Packages, 2nd Edition. O’Reilly Media, Inc.