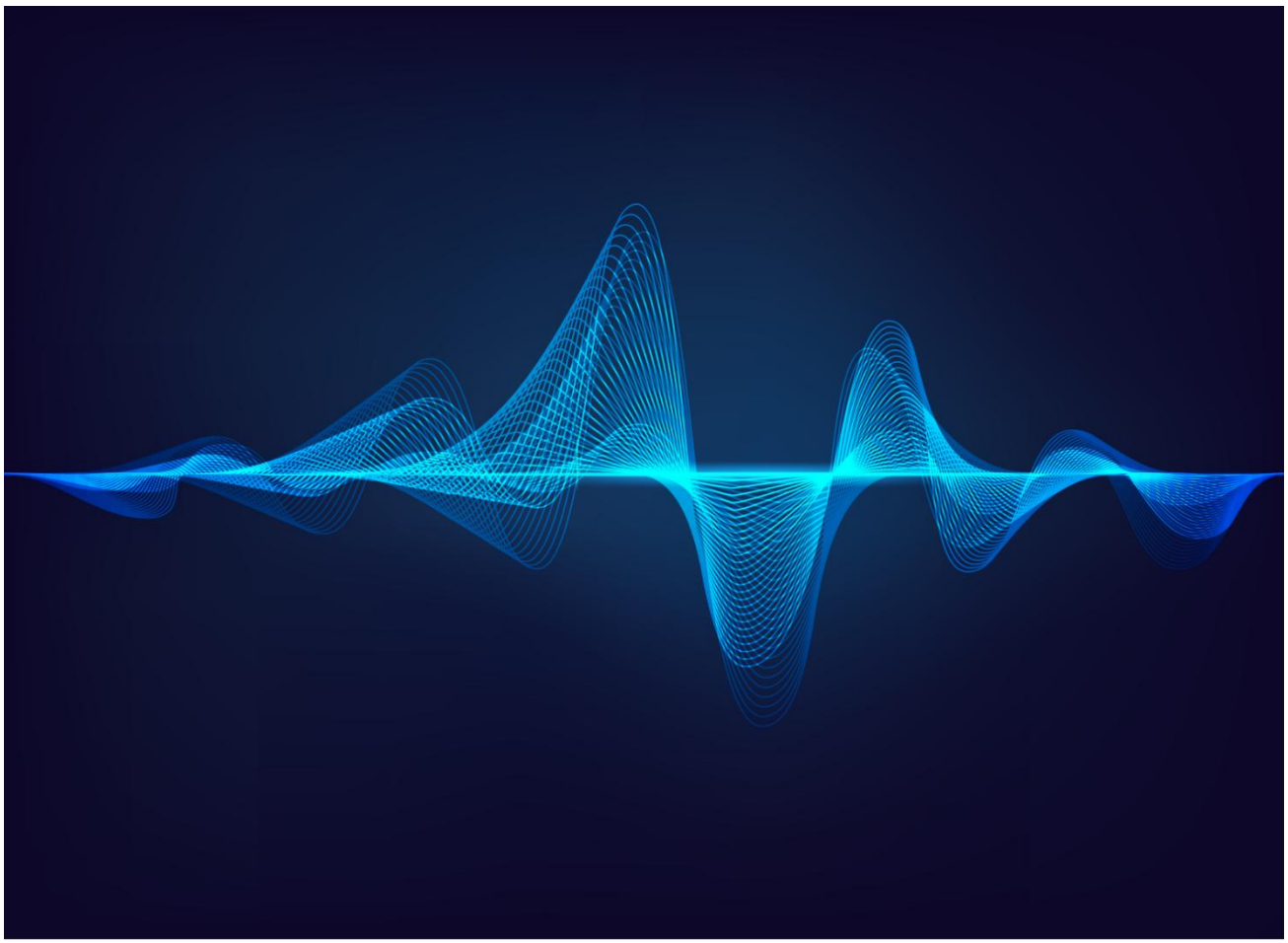


D4.4 – Survey Data Collection and Management Methods and Tools

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:	**D4.4**
Title:	**Survey Data Collection and Management Methods and Tools**
Lead beneficiary:	UTU
Work package:	WP4
Task:	T4.2
Dissemination level:	Public
Submission date:	30.07.2025
Revision/resubmission:	15.05.2026
Main author(s):	Aleksi Lahtinen (UTU)
Contributor(s):	UTU, SINUS, SSSA
Quality check:	Leo Lahti, UTU James Edwards, SINUS Arianna Martinelli, SSSA Olga Kavkaeva, SSSA

Document history:

Revision	Date	Main modification	Author
0.6	24/07/2025	Report draft	Aleksi Lahtinen
0.7	25/07/2025	Quality check	James Edwards
0.8	26/07/2025	Quality check	Arianna Martinelli, Olga Kavkaeva, Leo Lahti
0.9	29/07/2025	Final version	Aleksi Lahtinen
1.0	30/07/2025	Submitted to EC	SINUS
2.0	15/05/2026	Revisions in response to Review Consolidated Report	Aleksi Lahtinen, James Edwards

© 2025 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295

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

Bayesian statistics: A statistical approach that combines prior beliefs with observed data to estimate full distributions over model parameters.
Cumulative logistic regression: A statistical model for analysing ordinal outcome variables, estimating the probability of responses falling into ordered categories.
Dashboard: An interactive tool that visualizes data, allowing users to explore information dynamically through visual elements like charts and summaries.
GitHub repository: An online storage space where project materials (such as code, tutorials, and documentation) are hosted and shared.
R: A programming language and software environment widely used for statistical computing, data analysis, and visualization. It provides a vast ecosystem of packages and tools for tasks ranging from basic data manipulation to advanced statistical modelling and reporting.
R package: A modular collection of functions, data and documentation designed to extend the functionality of the R programming language.
Shiny: A R package for building interactive web applications directly from R, enabling dynamic user interfaces without needing web development knowledge.
shinyapps.io: A cloud-based hosting service for Shiny applications, allowing them to be accessed via the internet.
Ordinal variable: A categorical variable where the categories have a meaningful order (e.g., satisfaction levels) but the intervals between them are not necessarily equal.
Posterior distribution: The updated probability distribution of a parameter after incorporating both the prior information and observed data, using Bayes' theorem.
Prior information: Knowledge or assumptions about a parameter or system before observing data, usually expressed as a prior distribution in Bayesian analysis.
Probabilistic methods: Analytical techniques that incorporate uncertainty into estimates by modelling distributions rather than single-point values.
Quarto: An open-source scientific and technical publishing system, used here to create tutorials and documents combining code and analysis.

© 2025 OpenMusE | HORIZON-CL2-2022-HERITAGE-01-05 | Grant Agreement No. 101095295

Table of Contents

Executive Summary 7
Introduction 8
1 Deliverables 9
2 Conclusion 11
References 12

Executive Summary

Task T4.2 focuses on enhancing the accessibility, analysis, and dissemination of complex survey data through the development of interactive tools and advanced statistical methods. In support of this objective, two interactive dashboards were created using R and the Shiny framework to present data from the Live Music Census and the Cultural Access and Participation survey. These dashboards are designed for both technical and non-technical audiences, offering intuitive visualizations, such as pie charts and bar plots, and interactive features that allow users to filter data by variables like city and gender.

To address the analytical challenges posed by small sample sizes in subgroup-level data, Bayesian probabilistic cumulative logistic regression models were applied using the *brms* package in R. To support broader adoption a series of tutorials and reproducible workflows were published in both Quarto and PDF formats. These resources are openly available on GitHub and provide practical guidance for analysts working with ordinal data. The authors will support other OpenMusE partners in applying the methods developed to the final survey datasets, where appropriate.

In addition, the *eurostat* R package was significantly updated as part of this task. Version 4.0.0 was released to align with Eurostat's migration to the SDMX API, replacing deprecated services and improving the package's performance and interoperability. Key enhancements include multilingual metadata support, improved caching, clearer data provenance, and tighter integration with external tools. A tutorial demonstrating the use of the updated package for retrieving culture- and music-related data is also available on GitHub. This tutorial will support the OpenMusE researchers in defining the potential scope of the music sector / industry as represented in current EU statistical systems.

Together, these developments contribute to the infrastructure for survey data analysis and dissemination within the project. By improving data accessibility, enhancing the robustness of analytical approaches, and aligning with open science practices, they help advance the broader objectives of Task T4.2. These tools and resources are intended to support ongoing work across the OpenMusE project and provide a foundation for future collaboration and development.

The outputs are available at:

- Survey data dashboard workflows: https://github.com/openmusiceurope/T4.2_dashboards
- Ordinal survey data analysis workflows: https://github.com/openmusiceurope/T4.2_ordinal_survey_data_analysis
- Eurostat R tools: https://github.com/openmusiceurope/T4.1_eurostat
- Eurostat R Shiny browser: https://github.com/openmusiceurope/T4.1_shiny

A one-stop overview of how the OpenMusE datasets and software can be utilized to replicate its outputs – the "OpenMusE cookbook" – is available at: https://github.com/openmusiceurope/.github

Introduction

The OpenMusE project aims to provide open-source tools for managing survey data. Work Package 4 (WP4) focuses specifically on the collection, management, and analysis of statistical and survey data. This includes the development of tools that support both data exploration and advanced analysis.

Deliverable D4.4 is the outcome of Task 4.2 and introduces two interactive dashboards based on survey data collected within the OpenMusE project. These dashboards allow users to explore the data online without needing to process or manage it themselves. In addition, the deliverable offers analytical workflows for probabilistic analysis of ordinal survey data.

Task 4.2 emphasizes usability, interoperability, and methodological transparency in survey data handling. In line with these goals, D4.4 delivers three core outputs:

- Interactive dashboards for exploring survey data from the Live Music Census and cultural access and participation survey.
- A series of tutorials and workflows for conducting probabilistic ordinal data analysis using R.
- Updates to the *eurostat* R package that is used to retrieve data from the Eurostat API.

The dashboard, built with R Shiny, enables users, including non-specialists, to explore the survey data through interactive visualizations. The accompanying tutorials guide users through probabilistic modelling techniques suited for sparse ordinal survey data. Additionally, the updates done to the *eurostat* R package enable retrieving several datasets related to music and culture.

1 Deliverables

Data dashboards

As part of Task T4.2, interactive data dashboards were developed using R and the Shiny web application framework (Chang et al., 2025) to improve the accessibility and dissemination of survey data. These dashboards provide intuitive, web-based tools for exploring response patterns, making demographic comparisons, and identifying trends without requiring advanced technical skills. Visualizations, created using the *ggplot2* package (Wickham, 2016), include, for example, bar charts and pie charts that present both visual and numerical insights. Two dashboards were created: one for the Live Music Census[1], enabling comparisons across cities and gender for audience and venue survey responses, and another for the Cultural Access and Participation survey[2], covering cultural habits and political opinions across five European countries with filtering by country and gender. Both dashboards are hosted online and structured for flexible, user-friendly interaction tailored to different stakeholder needs. These versions of the dashboards are preliminary, as usability testing will be conducted later, and the dashboards will be updated based on the feedback received.

These OpenMusE outputs go beyond normal survey dashboard standards insofar as the backend code not only visualizes cleaned-and-prepared survey data, but acts as a "pipeline" from the raw dataset through stages of cleaning and preparation (e.g., removal of bad responses, removal of outliers, assigning / imputing missing values, etc.) to end-user-oriented visualization. Moreover, the code merges survey data with Eurostat data on economic inputs and outputs (fetched using the *eurostat* package described below) in order to generate estimates of the gross-value-added by live music across several categories of expenditure: merchandise, transport, hospitality, and tickets. Taken as a whole, this represents a software ecosystem that can be used to assess the total economic value of live music in more granularity than was possible before.

The public-facing dashboards are available at:

- Live Music Census: https://www.openmuse.eu/live-music-census/
- Cultural Access and Participation: https://www.openmuse.eu/cultural-access-and-participation/

The datasets and R code used to create the dashboards are available at:
https://github.com/openmusiceurope/T4.2_dashboards

Probabilistic ordinal survey data analysis workflows

A major challenge in analyzing data from the Live Music Census and the Cultural Access and Participation survey is the sparsity of responses at the subgroup level, which can undermine the reliability of standard analytical methods. To address this, Task T4.2 utilizes Bayesian probabilistic modeling, specifically applied to cumulative logistic regression (Bürkner & Vuorre, 2019) for analyzing ordinal data such as satisfaction ratings and agreement levels. The probabilistic framework can better handle scenarios where the amount of data is sparse, while also enabling the fitting of more complex models.

[1] Live Music Census dashboard: https://allahtinen.shinyapps.io/Census_dashboard/

[2] Cultural Access and Participation dashboard: https://allahtinen.shinyapps.io/cultural_participation_dashboard/

The cumulative logistic regression model, commonly used for analyzing ordinal survey responses, models the data by introducing an unobserved continuous latent variable. Instead of predicting the ordinal outcome directly, the model estimates the cumulative probabilities of a response falling at or below each category threshold (Agresti, 2010). These thresholds divide the latent variable into segments corresponding to the ordered response categories. This approach respects the ordinal nature of the data without assuming equal spacing between categories, making it more appropriate than methods designed for continuous outcomes. The probabilistic framework extends this model by incorporating prior knowledge and estimating full posterior distributions rather than single point estimates (Gelman et al., 2013). This enables greater flexibility, more robust inference in the face of sparse data, and natural quantification of uncertainty.

To support adoption of this approach, a suite of tutorials was developed using the *brms* package (Bürkner, 2017) in R. The package enables fitting probabilistic regression models, including ordinal models. The tutorials guide users through data preparation, model fitting, and interpretation, from basics to more advanced models. These resources are publicly available in both Quarto and PDF formats via GitHub repository[3], promoting accessibility and aligning with best practices in research data management by improving the robustness of insights from sparse survey data.

The analysis workflow and accompanying conference proceeding (Lahtinen et al. 2025) are available at: https://github.com/openmusiceurope/T4.2_ordinal_survey_data_analysis

Eurostat R package update

The *eurostat* R package underwent a significant update in the OpenMusE project to maintain compatibility with Eurostat's updated infrastructure, notably transitioning to the new SDMX API after the deprecation of the old JSON and bulk download services. This update led to the release of version 4.0.0, which introduced several improvements: enhanced caching to reduce redundant data retrieval and improve performance, support for multilingual metadata (adding German and French alongside English), improved citation and data provenance features, and better integration with external tools through updated metadata and version tracking. The package also replaced legacy dependencies (e.g., moving from *httr* (Wickham, 2023) to the more modern *httr2* (Wickham, 2025)) and optimized large dataset handling by supporting *data.table* package (Barrett et al., 2025). Mapping functiosns were decoupled and offloaded to the *giscoR* package (Hernangómez, 2025) to allow lighter installations. The previously mentioned GitHub repository also includes tutorial how the package can be used to retrieve culture and music related data.

The datasets and R code used to create the dashboards are available at: https://github.com/openmusiceurope/T4.2_dashboards

The analysis workflow and accompanying conference proceeding (Lahtinen et al. 2025) are available at: https://github.com/openmusiceurope/T4.2_ordinal_survey_data_analysis

The Eurostat updates and cultural and music data vignette are available at:

- R tools: https://github.com/openmusiceurope/T4.1_eurostat
- R Shiny browser: https://github.com/openmusiceurope/T4.1_shiny

[3] GitHub repository: https://github.com/Allaht2/OpenMusE_WP4_UTU

2 Conclusion

Task T4.2 has successfully advanced the OpenMusE project's objectives by delivering tools that improve the accessibility, analysis, and dissemination of survey data. The development of interactive dashboards and probabilistic modelling tutorials reflects a strong alignment with the task goals, emphasizing usability, methodological rigor, and open dissemination. These outputs not only address current project needs but also establish a scalable foundation for future reuse and cross-project collaboration.

A one-stop overview of how the OpenMusE datasets and software can be utilized to replicate its outputs – the "OpenMusE cookbook" – is available at: https://github.com/openmusiceurope/.github

References

Agresti, A. (2010). Analysis of Ordinal Categorical Data. John Wiley & Sons. https://doi.org/10.1002/9780470594001

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B, Krylov I (2025). *data.table*: Extension of `data.frame, R package version 1.17.6, https://CRAN.R-project.org/package=data.table

Bürkner P (2017). *brms: An R Package for Bayesian Multilevel Models Using Stan*. Journal of Statistical Software, 80(1), 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

Bürkner P-C, Vuorre M. *Ordinal Regression Models in Psychology: A Tutorial*. Advances in Methods and Practices in Psychological Science. 2019;2(1):77-101. DOI: https://doi.org/10.1177/2515245918823199

Chang, W., Cheng, J., Allaire, J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B. (2025). *shiny: Web Application Framework for R*. R package version 1.10.0.9001, https://shiny.posit.co/

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Hernangómez D (2025). *giscoR*: Download Map Data from GISCO API – Eurostat, R package version 0.6.1, https://ropengov.github.io/giscoR/

Lahtinen, A., Edwards, J. R., Calmbach, M., Tautscher, I., & Lahti, L. (2025). Robust modelling of ordinal survey data using probabilistic programming. In R. Ros (Ed.), *Computational Humanities Research 2025* (pp. 607–624). Anthology of Computers and the Humanities. https://doi.org/10.63744/eCwMjQ976nWf

Wickham H (2025). *httr2*: Perform HTTP Requests and Process the Responses, R package version 1.1.2, https://CRAN.R-project.org/package=httr2

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 (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org