Publishing and studying historical opera and music theatre performances on the Semantic Web: case OperaSampo 1830–1960

Annastiina Ahola[1], Eero Hyvönen[1,2], Heikki Rantala[1] and Anne Kauppala[3]

[1]*Semantic Computing Research Group (SeCo), Aalto University, Finland*
https://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[2]*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*
[3]*Sibelius Academy, University of the Arts, Finland*

Abstract

The OperaSampo is a Linked Open Data (LOD) service and semantic portal for searching, browsing, and analyzing information related to opera and music theatre performances performed in Finland during 1830–1960. The key data originates from the Reprises database of the Sibelius Academy, Finland. This paper presents the conversion process of transforming the data into LOD, the data model created, as well as the portal developed for using the data. The novelty of OperaSampo lays on its focus on using and studying data about the musical performances and persons involved in different roles. The system is available for final testing on the Semantic Web and will be published for open use in September 2023.

Keywords

Cultural Heritage, Linked Data, User Interfaces, Portals

1. Introduction

Reprises[1] is a database owned and maintained by the Sibelius Academy of Uniarts, Helsinki, Finland. The database has information on over 9000 opera, operetta, vaudeville, and other forms of music theatre performances that were performed in Finland during 1830–1960, involving over 3500 people in different roles. Currently the data covers the performances of the Esplanade Theater in Helsinki (1827–1857), the Swedish Theatre in Helsinki (1860—1961), the Russian Theatre in Helsinki (1868—1917), the Finnish Opera Company (ca. 1873—1879), and music theatre performances in the Turku Play-house (1839–1897). The data has been mainly sourced from theatre posters, newspaper advertisements, and repertoire books [1, 2, 3, 4, 5]. The eventual aim is to include all opera and music theatre performances that took place in Finland before 1960 except for those given by the Finnish National Opera[2], including also appearances made by visiting theatre troupes in Finnish cities.

0009-0008-6369-4712 (A. Ahola); 0000-0003-1695-5840 (E. Hyvönen); 0000-0002-4716-6564 (H. Rantala); 0000-0003-4113-1384 (A. Kauppala)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Available at: http://reprises.uniarts.fi/en/ (no longer accessible on up-to-date modern browsers)
[2]These are already available at the Encore service: http://encore.opera.fi

Reprises database has been used for both research and educational purposes at the University of the Arts, Finland. The traditional Reprises user interface, however, is not able to bring the full potential of the data into service for, e.g., Digital Humanities (DH) research [6]. Furthermore, the web application no longer works on up-to-date browsers since the spring of 2023 due to problems with updating the legacy software it is running on. As a solution, it was decided to create a new data service and portal based on semantic web technologies [7, 8] for publishing and using Cultural Heritage (CH) linked data [9].

The OperaSampo Knowledge Graph (KG) and OperaSampo Portal are the results of the work of transforming the data into LD form as well as of providing the end-user with new ways to search, browse, and analyse the data. The OperaSampo Portal was developed with the Sampo-UI framework [10, 11] based on the "Sampo model" [12] for CH data creation and publishing. The six principles of the Sampo model are listed in Table 1. The Sampo-UI framework facilitates the easy creation of semantic portals offering faceted search capabilities as well as integrated data-analytic tools for the user.

This paper presents the created KG and the semantic portal built on top of it. First, the KG and its creation are explored in Section 2. Then, the portal is presented in Section 3. Lastly, the related works are discussed and contributions of the paper summarized.

The portal is currently available for the final testing phase[3] and will be published for open use in September 2023. The data will be available using the CC BY 4.0 license on the Linked Data Finland platform LDF.fi [13] with an open SPARQL endpoint, schema documentation using the LODE service[4] [14]), dereferencing of URIs for both human users and machines, an RDF browser for technical users. The data will also be available as a data dump on the Zenodo repository[5].

Table 1
Sampo Model Principles P1–P6 for LOD service creation and portal design [12]

P1	Support collaborative data creation and publishing
P2	Use a shared open ontology infrastructure
P3	Make clear distinction between the LOD service and the user interface (UI)
P4	Provide multiple perspectives to the same data
P5	Standardize portal usage by a simple filter-analyze two-step cycle
P6	Support data analysis and knowledge discovery in addition to data exploration

2. OperaSampo Knowledge Graph

Conversion process The data from Reprises was exported as multiple CSV files for the conversion process to RDF with each CSV file having been a table in the original database. In total there were 16 relevant CSV files in the data dumps that were used.

The conversion was done with Python scripts requiring just the original CSV files initially. In this way the process could easily be run again whenever a new data dump was available. In

[3] operasampo.demo.seco.cs.aalto.fi/
[4] https://essepuntato.it/lode/
[5] https://zenodo.org/

the script the creation and serialization of the graph was done with the RDFLib[6]. The pandas[7] library was used for handling the CSV files and the tables read from them. The script goes over all the rows in the CSV files one by one using the IDs present in the CSV files to form the URIs in the final KG. Some of the fields inside the files used XML and some HTML-style formatting for the text, so some text handling was done with the libraries untangle[8] and BeautifulSoup[9].

Nine out of the 16 original CSV files corresponded well to class concepts – e.g., *performances* and *compositions* – and were be turned into classes in the OperaSampo KG. The rest of the files are instead turned into properties and their values for the aforementioned classes, e.g. translator information for performances.

One of the crucial parts of the data were the names of the performances as they were performed, reflecting the language of the performance, as opposed to the original title of the piece. This information was included in the Finnish texts in the additional information field in the original database and was extracted from there to its own property in the RDF form. This information was then also used to form the preferred labels for the performances together with the performance date. This process also highlighted performances that were missing the name information so they could be added manually.

Data enrichment The data from the original Reprises database was then supplemented with pictures of performance posters and venues by creating additional CSV files with the mapping from the entity IDs to the image files names or URLs depending on whether the image was a local file one or not. For the images of performance venues, a new *PlaceImage* class was created while the poster images used the already existing *PerformanceImage* class that was created from the original CSV files. The images of performance venues were matched manually to the venue IDs due to the low amount of images and venues. For the performance posters, the images were first renamed after the performances present in the poster and their performance dates (local files) or catalogued in a CSV file with the URLs and performance dates (online images). A Python script was then run that attempted to connect poster images to their performances based on the names and dates of performances. In addition, the script produces a file with the images that could not be connected to any performance and whether there were performances that matched one of the criteria – either name or date – but not the other. This file was used to correct some mistakes in the names of the performances in the data as well as highlighting some missing performances.

The information about the people mentioned in the data, but who were not actually performing, such as composers, was somewhat limited due to it not being the focus of the original database. To enrich the data shown for people in the portal in the future, the people are being linked to other information sources, such as the the KG of BiographySampo – Finnish biographies on the Semantic Web [15, 16], and the Wikidata KG.

Data model Figure 1 illustrates the general data model structure with an example performance entity in the OperaSampo KG. The entity represents a 'Trubaduri' performance of Verdi's composition 'Il trovatore' that was performed on the date 1874-04-12. The performance is linked to relevant performance-specific entities through properties has extra information in a textual

[6] https://github.com/RDFLib/rdflib
[7] https://pandas.pydata.org/
[8] https://untangle.readthedocs.io/en/latest/
[9] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

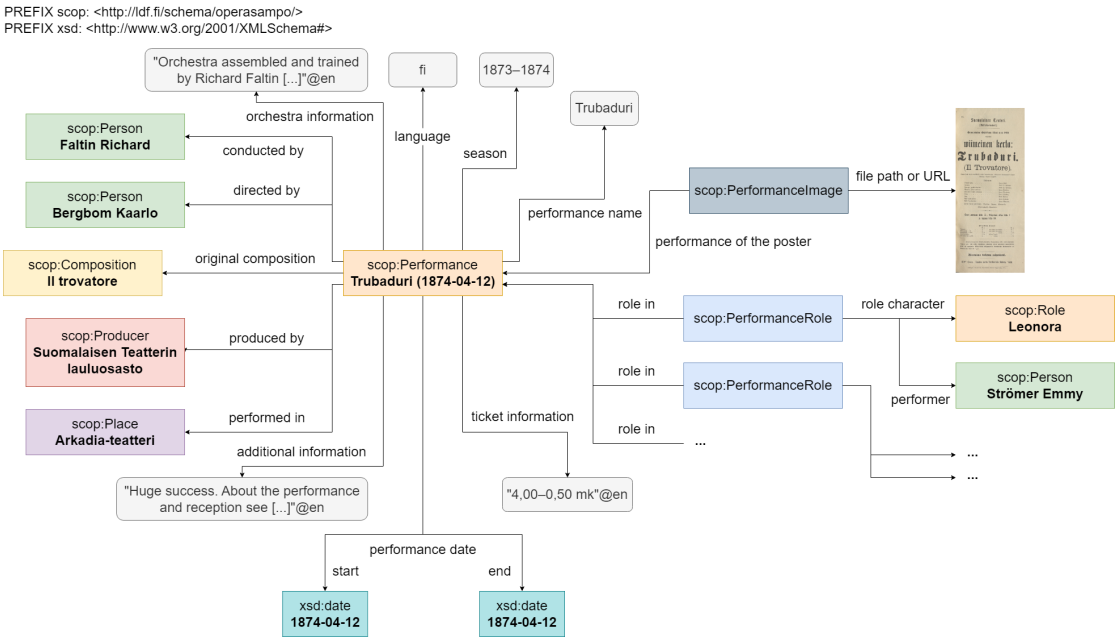

Figure 1: A chart illustrating the OperaSampo data model through an example.

format through properties with literal values.

Images of the posters – as well as the images of venues for performance venues – are linked from the *PerformanceImage* entity to the *Performance* entity or from *PlaceImage* to *Place* entity respectively. The image entities also hold potential description and copyright information if it is available for that particular image.

Roles for the performances are modeled through their own class, *PerformanceRole*, that has a property referring to the performance that it is a part of. This class connects a *Performance* entity to the performer of the role (*Person*) as well as the role character itself (*Role*) of a composition.

An important factor in designing the data model was to ensure that the data could still be easily edited by the same people in charge of the Reprises database. The final data structure largely follows the structure of different tables in the original database with few simplifications for the tables that could just be modeled as properties and their values instead. When possible, data types and formats were restricted, e.g., dates to xsd[10]:date, to try to ensure that a person inputting data in the future wouldn't be able to input data in an incorrect way. To deal with uncertainties regarding things like exact performance dates, the performance dates are inputted with both a start and end date even if the two dates are the same. This way the inputting of a new performance date is always the same process and doesn't, for example, require the creation of a new time span object. This could have been another way to model the dates.

Data service The data is stored on a Apache Jena Fuseki[11] SPARQL server with a Apache

[10]The 'xsd' prefix is used here for http://www.w3.org/2001/XMLSchema#
[11]https://jena.apache.org/documentation/fuseki2/

Lucene[12]-based text search. The data is accessible through an open SPARQL endpoint[13] that is also used by the OperaSampo Portal. In total, there are ca. 530,000 triples in the OperaSampo KG. The approximate instance counts for all the classes in the data are listed in Table 2.

Table 2
Instance counts for the classes in OperaSampo KG.

Class (Type)	Instances
Performances	9,000
People (e.g. authors)	3,500
Compositions	650
Role characters	6,400
Performance venues	60
Producers	150
Performance images	200
Place images	25

Data maintenance The aim is for the data to be managed in the LD format in the future instead of relying on the old database format, where it would have to be regularly converted using the same conversion scripts used for the original LD version of the data. The final choice of the platform used for editing the data is up to the data editors themselves with the Saha metadata editor [17] being one of the strong contenders.

3. User Interface

The OperaSampo Portal[14] is built with the Sampo-UI framework[15]. The contents of the portal are based on the wants and needs of the researchers that had been maintaining and using the original Reprises data to both offer an alternative to the old search interface of the Reprises database. GHe goal here was to enhance its search capabilities to a new level with the inclusion of faceted search and browsing [18, 19] with integrated data-analytic tools, as suggested in the Sampo model. In addition to the actual portal, also a video showcasing its basic functionality is also available on the Web[16].

Figure 2 depicts **the landing page** of the OperaSampo Portal. The UI is split into five *application perspectives* available as clickable rectangular cards on the screen:

1. *Performances*. This perspective covers all the performances in the data.
2. *People*. This perspective covers all the people related to performances and compositions, e.g., composers, translators, performers etc.
3. *Compositions*. This perspective covers all the compositions in the data.

[12]https://lucene.apache.org/
[13]https://ldf.fi/operasampo/sparql
[14]Source code available at: https://github.com/SemanticComputing/operasampo-web-app
[15]Source code available at: https://github.com/SemanticComputing/sampo-ui
[16]https://vimeo.com/805493196

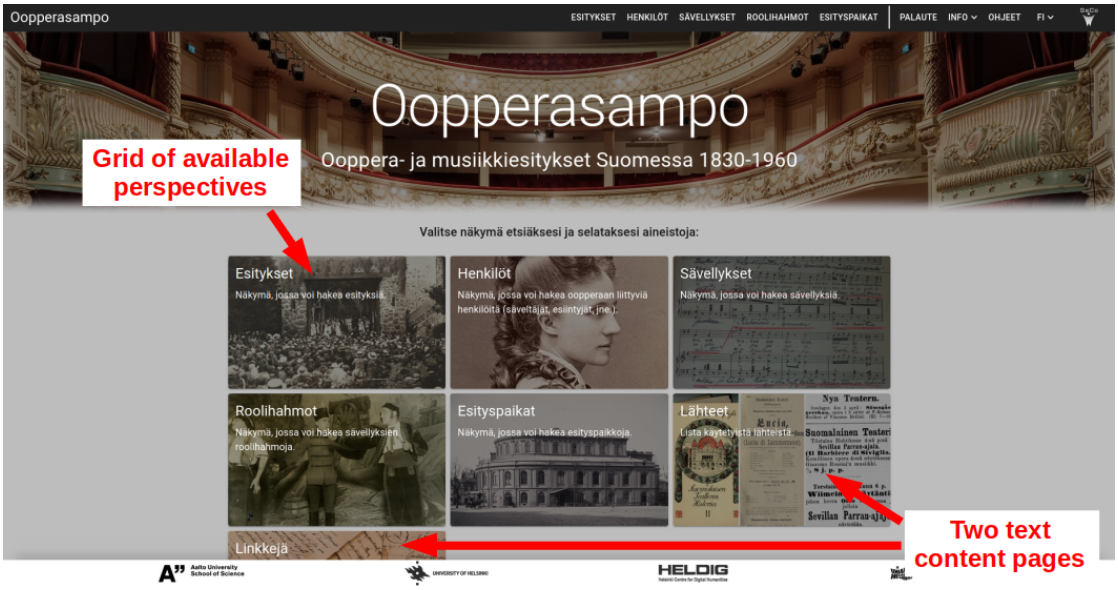

Figure 2: The landing page of the OperaSampo Portal.

4. *Role characters.* This perspective covers all the different role characters from compositions.

5. *Performance venues.* This perspective covers all the venues at which the performances in the data have been performed.

In addition to these perspectives, the landing page has two additional cards, *Sources* and *Links*. These two cards lead to pages with text content, either covering the sources used for the original data or links to other archives, respectively.

The faceted search view of the *Performances* perspective is shown in Figure 3. The user has made the selection *Verdi Giuseppe* in the *Composer* facet, which has filtered the results to only performances of compositions that have been composed by Verdi. Selecting any additional values from the *Composer* facet would update the results to include all performances of compositions composed by *either* of the selected composers. If the user were to select a value from another facet, the results would be filtered to fulfill the criteria of both of the facets. For example, if the user instead also selected *Arkadia-teatteri* from the *Performance venue* facet, the results would only include performances of compositions composed by Verdi *and* were performed in the Arkadia Theatre.

Clicking on underlined text in the table view leads to **the instance page** of that particular entity. If the user, for example, wants to see the detailed information on a particular performance, they can click the name of that performance to open up its instance page. Figure 4 shows the instance page for a performance of Verdi's *Il trovatore* performed on April 12th in 1874 with the name *Trubaduri*. The instance pages include at least the same information on that entity as what is present in the table view, but can also include additional information that wasn't included as a column in the table.

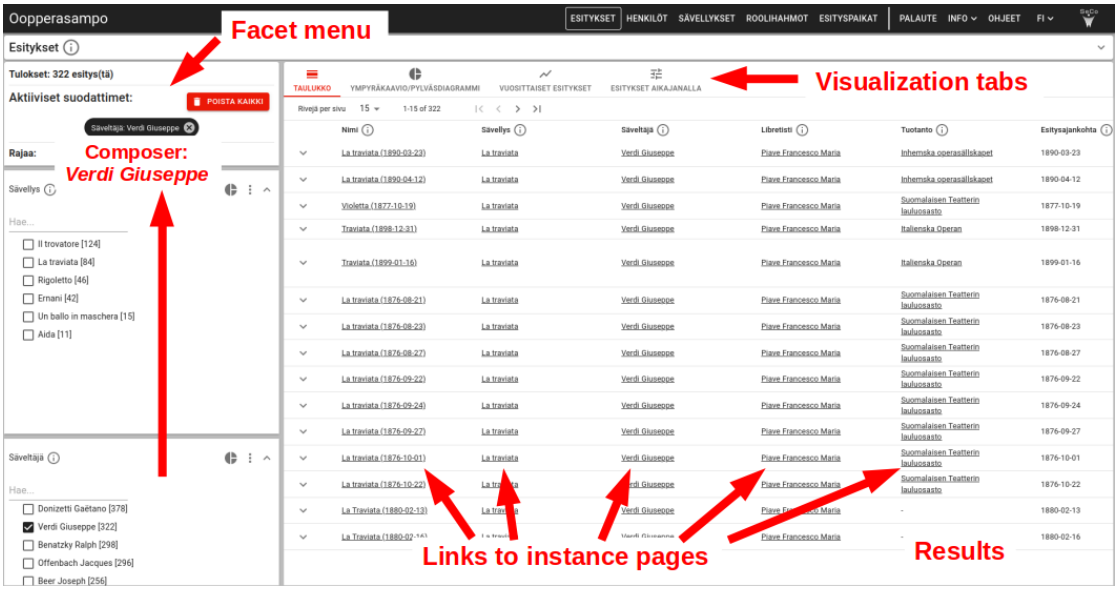

Figure 3: The faceted search view of the Performances perspective.

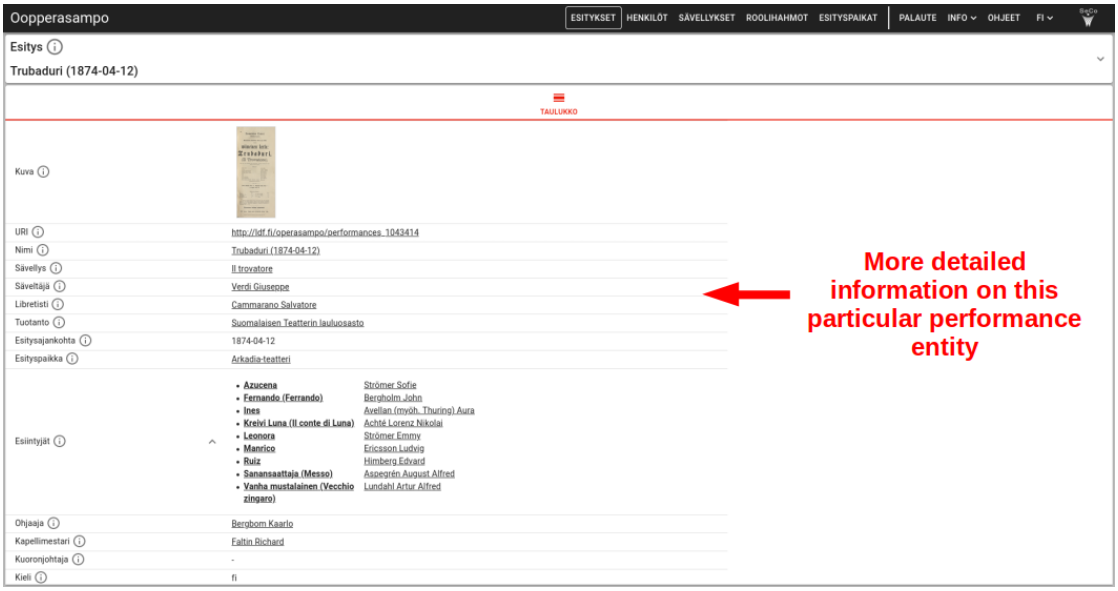

Figure 4: The instance page of a performance entity

In addition to the table view of the faceted search view, there are various **visualization tabs** available in the different application perspectives. Figure 5 shows a timeline visualization tab included in the *Performances* perspective, where performances are visualized on a timeline in

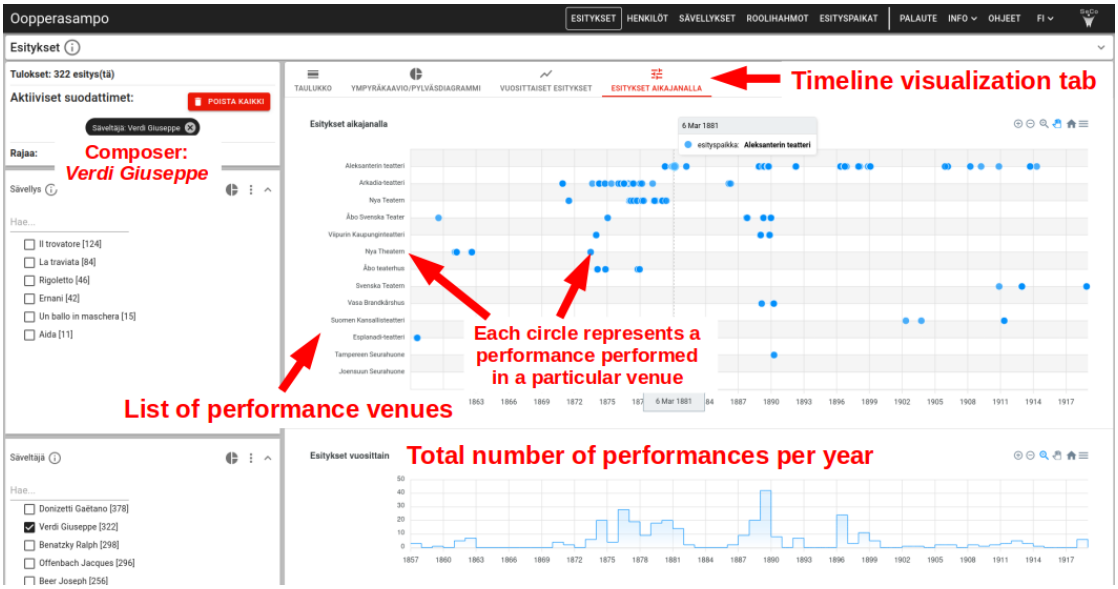

Figure 5: A visualization illustrating the annual number of performances as well as the venues they were performed in

respect to the venue where they were performed in. This particular visualization, for example, enables the user to gauge the reception of certain performance subsets, e.g., how often and when Verdi's works were being performed and whether there were certain theatres where they were performed in more often than others.

Other particularly interesting part of the data from the perspective of the DH researchers using the Reprises data for research are the performers themselves as well as the roles that they were performing during their career and their chronology. In the *People* perspective, each person entity has a column for the possible role characters they have performed. The instance pages in the people perspective also include an additional row for a list of performances that particular person has performed in. The performances in that row are grouped by the role character that a particular person performed as in that performance. Figure 6 showcases how the list of role characters and performances are found for the opera singer *Wäinö Sola*. The names of the performances also include the performance date for them, so the user can gauge the evolution of the performer's voice throughout the years based on the role characters they have performed as and their voice types.

In the *Role characters* perspective the user can explore this same information from the opposite perspective, that is, which performers have performed as a certain role character. Figure 7 illustrates an example case where the user wants to see the list of people who have performed in the role *Il conte di Luna* (Verdi: *Il trovatore*).

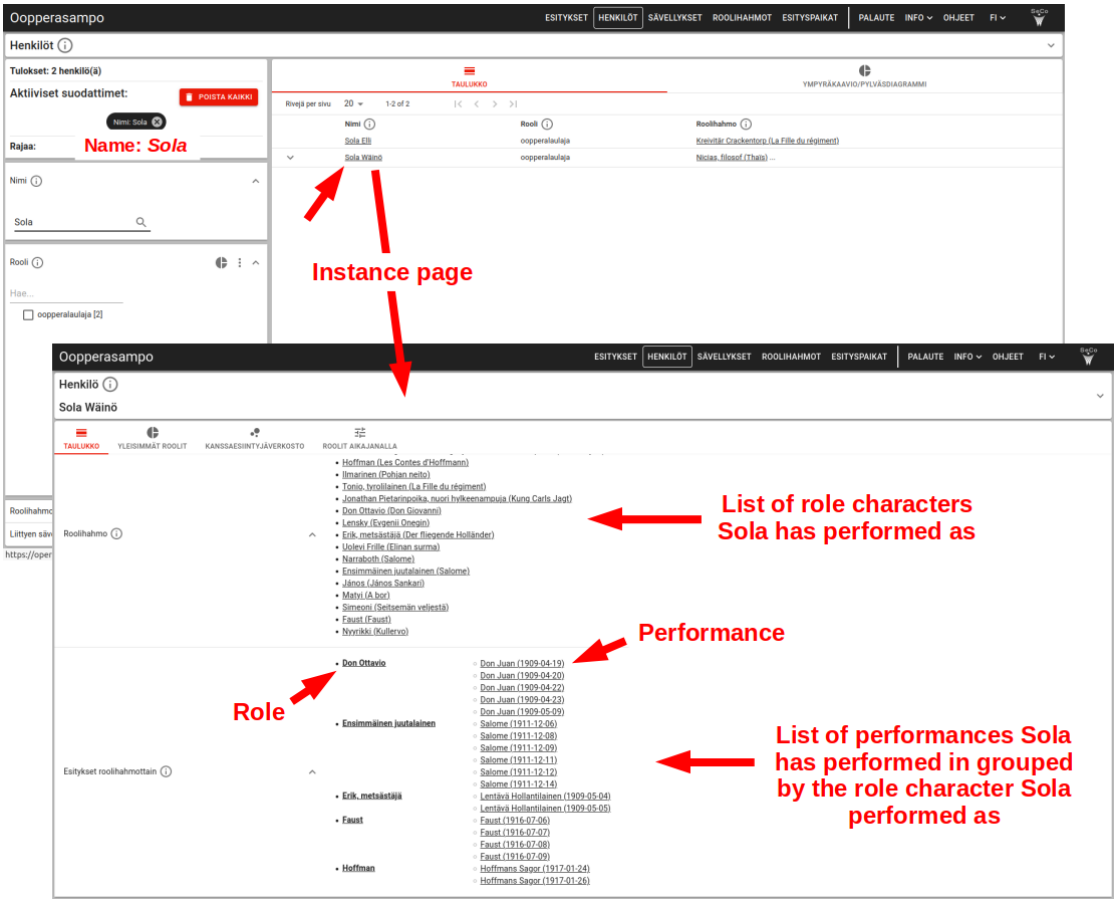

Figure 6: The instance pages of performers have a list of the role characters the performer has performed, as well as a list for the performances the performer has performed in grouped by the role character.

4. Related work and Discussion

A multitude of opera and theatre archives exist – such as Svenska operans repertoararkiv[17]; Dansk Forfatterleksikon[18]; The London Stage Database, 1660–1800[19]; Archives de l'Opéra Comique[20]; Les Archives du Spectacle[21]; Staatsoper Dresden database[22]; Operatic Productions in the Nertherlands 1886–1995[23] and Music in the Second Empire Theatre[24] – but the vast majority of them to our knowledge do not use LD. An exception to these is a dataset from the

[17] https://arkivet.operan.se/repertoar/
[18] http://danskforfatterleksikon.dk/1850t/t1850t.htm
[19] https://www.eighteenthcenturydrama.amdigital.co.uk/LondonStage/Database
[20] https://dezede.org/dossiers/archives-opera-comique/data
[21] https://www.lesarchivesduspectacle.net
[22] http://test.performance.slub-dresden.de/projects/staatsoper-dresden
[23] https://brill.com/downloadpdf/journals/rdj/5/2/article-p79_79.pdf
[24] http://www.fmc.ac.uk/mitset/index.html#/

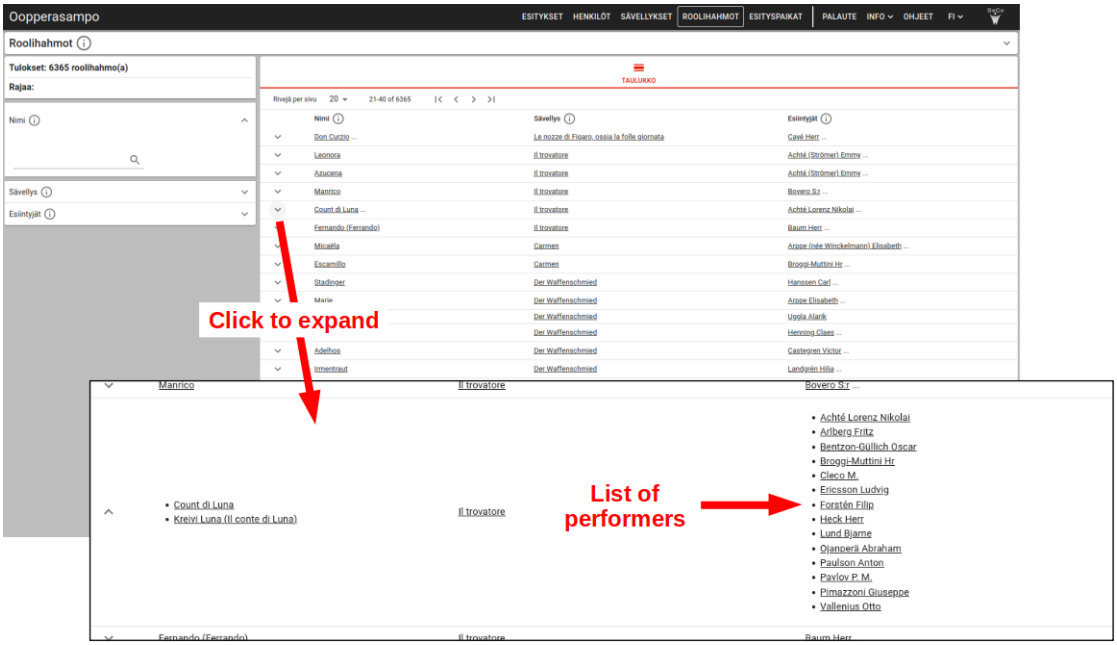

Figure 7: A list of performers that have performed as that particular role character is included for every role character entity.

Stuttgart State Theatres that was turned into a browsable LD KG, the Linked Stage Graph [20], as a part of the Coding da Vinci initiative CH hackathon.

In the field of music, LD has been used, for example, to model relationships between jazz musicians in the Linked Jazz[25] system [21]. Various metadata archives, such as the open music encyclopedia MusicBrainz[26], Wikidata, and Live Music Archive[27], also exist. The Live Music Archive has been turned into LD form [22] and has been used for audio analysis of live music performances in [23] and [24]. Using LD in representing the relation between performances and scores is discussed in [25], while in [23] LD is used for finding performance recordings in an archive. In contrast, the novelty in OperaSampo lays on its focus on using studying historical data about the musical performances and persons involved in different roles.

The OperaSampo Portal offers a novel way of exploring and searching performance data in a way that wasn't possible before. Faceted search and browsing gives the user the possibility to both set very specific filters on the results as well as to explore the data in a more iterative way to find interesting data without any predetermined idea of what it might be. With the integrated data-analytic tools, the user can easily perform basic analyses and visualization on the data without having to manually count or calculate the results.

[25] https://linkedjazz.org/
[26] https://musicbrainz.org/
[27] https://archive.org/details/etree

Acknowledgments

Thanks to Outi Parkkila, Anne Piirainen and others at the Sibelius Academy and University of the Arts for their collaboration on this project. Computational resources of CSC – IT Center for Science were used.

References

[1] E. Aspelin-Haapkylä, Suomalaisen Teatterin historia, I–IV, Finnish LIterature Society SKS, Helsinki, 1906–1909.

[2] L. Byckling, Keisarinajan kulisseissa: Helsingin venäläisen teatterin historia 1868–1918, Finnish Literature Society SKS, Helsinki, 2009.

[3] M. Lüchou, Svenska teatern i Helsingfors: repertoar, styrelser och teaterchefer, konstnärlig personal 1860–1975, Stiftelsen för Svenska teatern i Helsingfors/Söderström, Helsinki, 1977.

[4] M. M. van Nieuwkerk, L. Salters, R. M. Helmers, I. Kisjes, Operatic productions in the Netherlands, 1886–1995: from printed annals to searchable performance data, Research Data Journal for the Humanities and Social Sciences 5 (2020) 79–90. URL: https://brill.com/view/journals/rdj/5/2/article-p79_79.xml?language=en.

[5] P. Paavolainen, Arkadian arki: Kaarlo Bergbomin elämä ja työ, II, 1872–1887, Teatterikorkeakoulun julkaisusarja 51, Taideyliopiston Teatterikorkeakoulu, Helsinki, 2016. URL: https://urn.fi/URN:ISBN:978-952-6670-86-7.

[6] E. Gardiner, R. G. Musto, The Digital Humanities: A Primer for Students and Scholars, Cambridge University Press, New York, NY, USA, 2015. https://doi.org/10.1017/CBO9781139003865.

[7] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition), Morgan & Claypool, Palo Alto, California, 2011. URL: http://linkeddatabook.com/editions/1.0/.

[8] P. Hitzler, M. Krötzsch, S. Rudolph, Foundations of Semantic Web technologies, Springer, 2010.

[9] E. Hyvönen, Publishing and using cultural heritage linked data on the Semantic Web, Morgan & Claypool, Palo Alto, California, 2012.

[10] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, Semantic Web – Interoperability, Usability, Applicability 13 (2022) 69–84. doi:10.3233/SW-210428.

[11] H. Rantala, A. Ahola, E. Ikkala, E. Hyvönen, How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework, 2013. Paper submitted for VOILA 2023 workshop for evaluation.

[12] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability 14 (2023) 729–744. doi:10.3233/SW-223034.

[13] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: ESWC 2014: The Se-

mantic Web: ESWC 2014 Satellite Events, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[14] S. Peroni, D. Shotton, F. Vitali, The live OWL documentation environment: A tool for the automatic generation of ontology documentation, in: A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, N. Aussenac-Gilles, N. Hernandez (Eds.), Knowledge Engineering and Knowledge Management. EKAW 2012, Springer, Berlin, Heidelberg, 2012, pp. 398–412. doi:10.1007/978-3-642-33876-2_35.

[15] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – Publishing and enriching biographies on the Semantic Web for digital humanities research, in: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019), Springer, 2019, pp. 574–589.

[16] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case national biography of finland, Semantic Web – Interoperability, Usability, Applicability 14 (2023) 385–419. URL: https://doi.org/10.3233/SW-222887.

[17] J. Kurki, E. Hyvönen, Collaborative metadata editor integrated with ontology services and faceted portals, in: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010) at ESWC 2010, volume 596, CEUR Workshop Proceedings, http://CEUR-WS.org, 2010. URL: https://ceur-ws.org/Vol-596/paper-02.pdf.

[18] M. Hearst, Design recommendations for hierarchical faceted search interfaces, in: ACM SIGIR workshop on faceted search, Seattle, WA, 2006, pp. 1–5.

[19] D. Tunkelang, Faceted search, Morgan & Claypool Publishers, CA, USA, 2009.

[20] T. Tietz, J. Waitelonis, K. Zhou, P. Felgentreff, N. Meyer, A. Weber, H. Sack, Linked stage graph, in: SEMANTICS Posters&Demos, volume 2451, CEUR Workshop Proceedings, http://CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2451/paper-27.pdf.

[21] M. C. Pattuelli, K. Hwang, M. Miller, Accidental discovery, intentional inquiry: Leveraging linked data to uncover the women of jazz, Digital Scholarship in the Humanities 32 (2016) 918–924. URL: https://doi.org/10.1093/llc/fqw047. doi:10.1093/llc/fqw047.

[22] S. Bechhofer, K. Page, D. De Roure, Hello cleveland! linked data publication of live music archives, in: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), IEEE, 2013, pp. 1–4.

[23] S. Bechhofer, K. R. Page, D. M. Weigl, G. Fazekas, T. Wilmering, Linked data publication of live music archives and analyses, in: The Semantic Web – ISWC 2017, 2017, pp. 29–37. doi:10.1007/978-3-319-68204-4_3.

[24] K. R. Page, S. Bechhofer, G. Fazekas, D. M. Weigl, T. Wilmering, Realising a layered digital library: exploration and analysis of the live music archive through linked data, in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2017, pp. 1–10.

[25] J. Devaney, H. L. Gauvin, Representing and linking music performance data with score information, in: Proceedings of the 3rd International Workshop on Digital Libraries for Musicology, DLfM 2016, Association for Computing Machinery, New York, NY, USA, 2016, p. 1–8. URL: https://doi.org/10.1145/2970044.2970052. doi:10.1145/2970044.2970052.