*Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI*
*A. Salatino et al. (Eds.)*

*doi:10.3233/SSW240031*

# Historical Opera and Music Theatre Performances on the Semantic Web: OperaSampo 1830–1960

Annastiina AHOLA [a], [1], Eero HYVÖNEN [a,b], Heikki RANTALA [a], and Anne KAUPPALA [c]

[a] *Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*
[b] *University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland*
[c] *Sibelius Academy, University of the Arts, Finland*

ORCiD ID: Annastiina Ahola https://orcid.org/0009-0008-6369-4712, Eero Hyvönen https://orcid.org/0000-0003-1695-5840, Heikki Rantala https://orcid.org/0000-0002-4716-6564, Anne Kauppala https://orcid.org/0000-0003-4113-1384

**Abstract.** The OPERASAMPO is a Linked Open Data (LOD) service and semantic portal for searching, browsing, and analyzing information related to historical 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 process of transforming the original data into LOD and the data model created for it, data maintenance, as well as the portal and data service for utilizing the data. The novelty of OPERASAMPO lays on its focus on studying data about the musical performances and persons involved in different roles using faceted search and browsing combined seamlessly with data-analytic tools for Digital Humanities research. The service was published for open use in October 2023.

**Keywords.** Cultural Heritage, Linked Data, User Interfaces, Portals

## 1. Introduction

According to a definition[2], *musical culture refers to the collective behaviors, practices, traditions, beliefs, and values related to music within a particular society or group. It encompasses various aspects such as musical genres, performance styles, instruments used, dance forms associated with music, and rituals surrounding music*. Digital Humanities (DH) methods [1] provide novel computational quantitative methods for studying cultural data.

This paper concerns DH research on one particular aspect of musical culture: historical music theatre performances in which musicians and other actors participate in

[2] https://library.fiveable.me/key-terms/ap-hug/musical-culture

different roles. A case study including a demonstrator is presented where data from a legacy database has been transformed into a Linked Open Data (LOD) knowledge graph (KG) in a data service including a SPARQL endpoint for research and application development. As a practical example of applications, a semantic portal OPERASAMPO in use on the Web is presented. This system extends the original legacy system by facilitating more flexible search and browsing, based on semantic faceted search, and integrated data-analytic tools for DH research. The OPERASAMPO PORTAL was developed with the Sampo-UI framework [2,3] based on the "Sampo model" [4] for CH data creation and publishing.

Both the portal[3] and the data set[4] have been published for open use in October 2023. The data is available under the CC BY 4.0 license on the Linked Data Finland platform with an open SPARQL endpoint[5].

This paper presents the created KG and the semantic portal built on top of it as well as evaluation on the functionality of the portal for researchers working with this kind of data. In the following, related works are first discussed in Section 2 to give context for our work, and the primary legacy data are introduced (Section 3). The data transformation process and the created KG is introduced in Section 4. Then, the portal and the data service are presented in Section 5. Section 6 goes over the maintenance of the KG. The evaluation of the portal is discussed in Section 7. Lastly, the contributions are discussed and summarized in Section 8.

This paper extends the our workshop [5] and short demo paper [6] about OPERASAMPO substantially by extending data model descriptions, giving new examples of using the data, discussion on maintaining the data, and by presenting an evaluation of the system.

## 2. Related Work

A multitude of digital opera and music theatre archives exist—such as Svenska operans repertoararkiv[6]; Dansk Forfatterleksikon[7]; The London Stage Database, 1660–1800[8]; Archives de l'Opéra Comique[9]; Les Archives du Spectacle[10]; Staatsoper Dresden database[11]; Operatic Productions in the Nertherlands 1886–1995[12] and Music in the Second Empire Theatre[13]—but the vast majority of them to our knowledge do not use LD. In addition, most of them are databases of a particular theatre and thus catalogue only their in-house performances. For instance, because Encore[14] is the Finnish National Opera and Ballet performance database, it does not contain any performances produced

[3] https://oopperasampo.fi/

[4] https://www.ldf.fi/dataset/operasampo

[5] http://ldf.fi/operasampo/sparql

[6] https://arkivet.operan.se/repertoar/

[7] http://danskforfatterleksikon.dk/1850t/t1850t.htm

[8] https://www.eighteenthcenturydrama.amdigital.co.uk/LondonStage/Database

[9] https://dezede.org/dossiers/archives-opera-comique/data

[10] https://www.lesarchivesduspectacle.net

[11] http://test.performance.slub-dresden.de/projects/staatsoper-dresden

[12] https://brill.com/downloadpdf/journals/rdj/5/2/article-p79_79.pdf

[13] http://www.fmc.ac.uk/mitset/index.html#/

[14] https://encore.opera.fi/en

by other company, not even when they have used its own stage. In order to have a thorough comprehension of the history of Finnish music theatre performances since 1910s, consultation of both Encore and OPERASAMPO is required. Unfortunately, presently their datasets cannot be linked, which limits the possibilities of accurate analysis.

An exception to these is a dataset from the Stuttgart State Theatres that was turned into a browsable LD KG, the Linked Stage Graph [7], as a part of the Coding da Vinci initiative CH hackathon. However, the Linked Stage Graph is not a performance database at all but a truly interesting way to view performance photographs and metadata related to them from Stuttgart Theatre (1890s–1940s) housed at the National Archive of Baden-Wuerttemberg. The collection of photographs in the Linked Stage Graph does not comprehensively cover all the performances in the time period. For instance, Richard Strauss's *Salome* was performed in the Stuttgart Theatre for the first time already in 1906 and between 1912 and 1942 it had several runs [8], but there is not a single photograph of these in the Linked Stage Graph, and therefore *Salome* remains non-existent in the database. This example is illustrative at a general level regarding the focus of the Linked Stage Graph.

In the field of music, LD has been used, for example, to model relationships between jazz musicians in the Linked Jazz[15] system [9] as well as between music history personalities—also including some of the more internationally well-known people present in the OPERASAMPO KG—in the Musical Meetups Knowledge Graph [10] as a part of the larger music heritage project Polifonia [11]. The DOREMUS [12] KG deal with classic music works and their performances. Various metadata archives, such as the open music encyclopedia MusicBrainz[16], Wikidata, and Live Music Archive[17], also exist. The Live Music Archive has been turned into LD form [13] and has been used for audio analysis of live music performances in [14] and [15]. Using LD in representing the relation between performances and scores is discussed in [16]. Historical data on Italian secular music and lyric poetry has been modeled using LD in the RePIM in LOD project [17].

## 3. Primary Data

OPERASAMPO is based on the *Reprises* database created and maintained by the Sibelius Academy, Finland, one of Europe's largest music academies. Reprises data includes information on over 9000 opera, operetta, vaudeville, and other forms of music theatre performances that were performed in Finland during 1830–1960. Over 3500 people have been involved in these performances in different roles. The database was created as research instrument for two research projects on history of opera and music theatre performances in Finland and Nordic countries to be publicly opened after the research projects. Besides performance scholars, the database project group included professionals of digital data management in Sibelius Academy.

The data covered the opera and music theatre performances of the major theatres, opera companies and play-houses listed in Table 1 during the specified time periods along with data from various smaller theatres. In addition to the premieres, also the suc-

[15] https://linkedjazz.org/
[16] https://musicbrainz.org/
[17] https://archive.org/details/etree

**Table 1.** Major theatres, opera companies and play-houses covered by the data

| Theatre, company or play-house | Time period |
|---|---|
| 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 |
| the Turku Play-house | 1839–1897 |

cessive performances and possible later runs were catalogued in order to be able to offer a reliable view of the historical performance culture. The data was mainly sourced from theatre posters, newspaper advertisements, and repertoire books [18,19,20,21,22]. The eventual aim was (and still 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[18], including also appearances made by visiting theatre troupes in Finnish cities.

The Reprises database has been used for both research and educational purposes at the University of the Arts, Finland. The traditional Reprises user interface, however, was not able to bring the full potential of the data into service for, e.g., Digital Humanities (DH) research. Furthermore, the web application ceased to work on up-to-date browsers during the spring of 2023 due to problems with the underlying legacy software. As a solution, it was decided to create a new data service and portal based on semantic web technologies [23,24] for publishing and using Cultural Heritage LD. [25].

## 4. OperaSampo Knowledge Graph: Transformation & Data Model

This section first introduces the transformation process done for the data to turn it into LD. Afterwards, the data model is introduced in more detail.

**Transformation** The data used for the transformation originates from a data dump of the Reprises database. The data dump was first converted to CSV files, where each file represented a single table in the original Reprises database. In total, 18 CSV files were created from the data dump; 16 out of which would be relevant for the transformation process.

The transformation from the CSV files to RDF format was done using Python scripts to also enable some needed data processing to be done directly during the conversion process. The Python scripts went over the relevant CSV files with the Pandas[19] library and a RDF graph was created out of them with the `rdflib`[20] library. The `Untangle`[21] and `BeautifulSoup`[22] libraries were used for handling some textual information present in the database's fields. The scripts were written in a way for them to be able to be reused with later data dumps during the transition period from the Reprises database to editable LD. The later data dumps were produced the same way as the first; a dump was first taken from the Reprises database and then converted into CSV files for the scripts. The scripts produced a serialization of the RDF graph in Turtle format.

[18] These are already available at the Encore service: http://encore.opera.fi

[19] https://pandas.pydata.org/

[20] https://rdflib.readthedocs.io/

[21] https://untangle.readthedocs.io/

[22] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

In the end, the conversion process had three main steps in total:

1. *Extracting names of performances from additional information fields.* The Reprises database didn't have a separate name field for performances. Instead, in the original Reprises database, the preferred names[23] of the performances were listed at the start of the additional information text field in bold. The `Untagle` and `BeautifulSoup` libraries were used for extracting these names—for parsing the text content from XML format or for getting bolded text respectively—to be used in forming the final preferred labels for performances (e.g., *Trubaduri (1874-04-12))*.
2. *Linking poster images to performances.* Some images of performance posters were compiled for the performances in the data. There were two types of images: (1) local files and (2) images available online. The local files were named according to the names and dates of the performances present in the posters while images available online were catalogued in a CSV file with the same name and date information connected to the URL of the image. These images were linked automatically to existing performance entities by matching the dates and names of the performances. This matching was updated to an existing CSV file for a table in the original Reprises database that was created for the purpose of linking images to performances to be used in the later `rdflib` conversion step.
3. *Conversion to RDF format with `rdflib`.* This step comprises of going over all the original and/or updated CSV files and adding the relevant information as triples into the created graph. Whenever information in the text fields was available in multiple languages, all versions of the text were added with the respective language tags attached.

Towards the end of the transition period from the original database to managing the data in LD format, an additional correction importing step was added before the final `rdflib` conversion step. This step was added to the conversion process after the editor for the Reprises database ceased to function on modern up-to-date browsers and additions and corrections could no longer be made there, but a newer data dump of the database could still be taken. This step simply automatically supplemented the existing CSV files with corrections and additions and created a new, updated version of the file to be used in the final conversion step.

Out of the 16 relevant CSV files, nine classes were created and the rest of the information was turned into appropriate properties and their values for the aforementioned created classes. These nine classes were supplemented with an additional class not present in the original database for the images of performance venues as well as additional properties based on the needs of the data editors and the UI for better search results (e.g., fields for additional labels for venues). A data model explained below was also defined for all the classes and properties in the OperaSampo data and included in the Turtle file.

**Data Model** The structure of the data model is largely based on the original data structure used in the Reprises database to ensure that the data could easily be edited by the

[23] Preferred names here refers to the names under which these performances were performed, e.g., *Trubaduri* instead of *Il trovatore*. These names were often also an indication of the language of the performance. Due to occasional data feeding errors, Reprises failed to recognize these linguistic variants as referring to the same opera composition, which caused uncertainty regarding the result set.

original data owners and/or editors even in LD format. The classes and their approximate instance counts in the data are listed in Table 2 while Figure 1 shows how different classes can connect to other classes. Figure 2 illustrates how a singular performance—*Trubaduri* (Verdi's *Il trovatore*) performed on April 12, 1874—is modeled in the KG.

**Table 2.** Rounded instance counts for all the classes (types) in the data

| Class (Type) | Count |
|---|---:|
| Performances | 9,000 |
| People (e.g., performers, composers, ...) | 3,500 |
| Compositions | 650 |
| Role characters | 6,500 |
| Performance venues | 60 |
| Producers | 150 |
| Performance poster images | 200 |
| Venue images | 50 |
| Occupation roles (e.g., 'opera singer') | 10 |
| Performance roles | 92,000 |

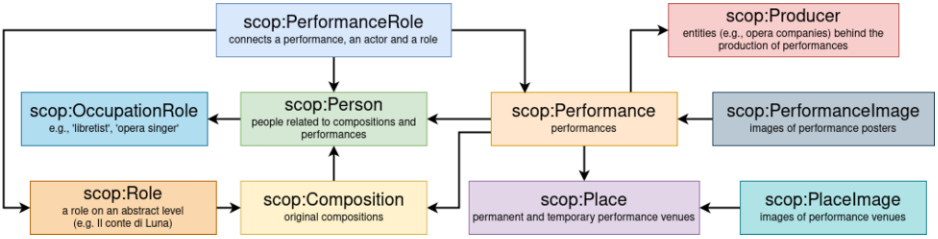

scop:PerformanceRole
connects a performance, an actor and a role
scop:Producer
entities (e.g., opera companies) behind the production of performances
scop:OccupationRole
e.g., 'libretist', 'opera singer'
scop:Person
people related to compositions and performances
scop:Performance
performances
scop:PerformanceImage
images of performance posters
scop:Role
a role on an abstract level
(e.g. Il conte di Luna)
scop:Composition
original compositions
scop:Place
permanent and temporary performance venues
scop:PlaceImage
images of performance venues

**Figure 1.** A chart illustrating how the different classes in OperaSampo KG connect to each other

The data model in OperaSampo defines the original works (compositions) and their manifestations (performances) on two levels: The `Composition` class represents the original compositions and the `Performance` class their manifestations, the performances performed. This is a simplification of the actual reality of what the exact content of a performance was. For example, in the case of Verdi's *Il trovatore*, though the data model links the performance straight to the original composition *Il trovatore*, the actual version performed is close also to the French version of the opera *Le Trouvère*, because the *mise-en-scène* of this French version was adapted to the Royal Swedish Theater's performances (*Trubaduren,* first performance 1860) whereas the libretto was translated to Swedish from the German adaptation of the Italian version [26]. An adaptation of this Swedish version was brought to Finland in 1861 and 1862, which served as a model for the 1870 Finnish language adaption of opera, without ballets (*Trubaduri*) [27] (see Figure 3). This was a conscious choice made in the original database design: The focus on the data was on the performances and performers themselves not in the original compositions, which in the performance practice of opera are not a self-value, and besides, in many cases challenging to determine. This focus is also apparent from looking at the compositions included in the data or the lack thereof—only compositions that have

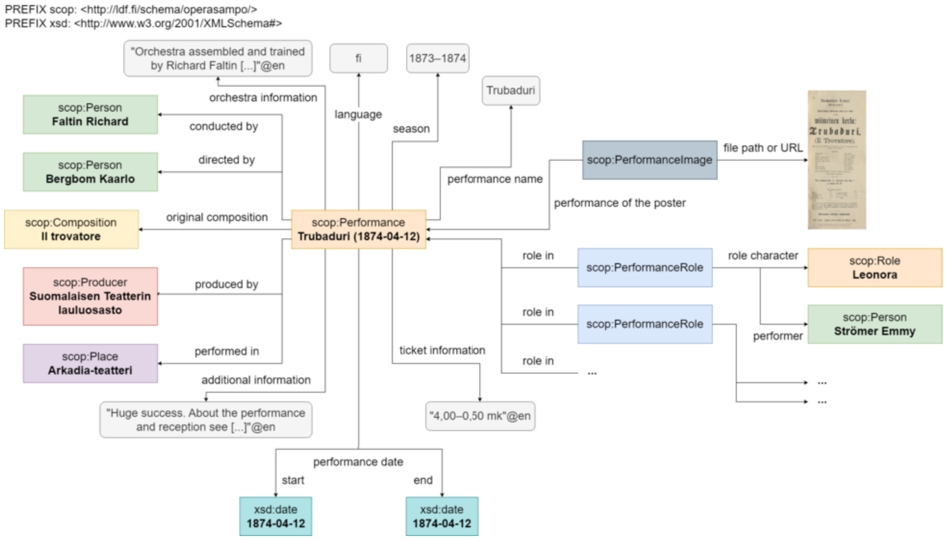

PREFIX scop: <http://ldf.fi/schema/operasampo/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
"Orchestra assembled and trained by Richard Faltin [...]"@en
fi
1873–1874
Trubaduri
orchestra information
language
season
scop:Person Faltin Richard
conducted by
scop:Person Bergbom Kaarlo
directed by
performance name
scop:PerformanceImage
file path or URL
performance of the poster
scop:Composition Il trovatore
original composition
scop:Performance Trubaduri (1874-04-12)
role in
scop:PerformanceRole
role character
scop:Role Leonora
scop:Producer Suomalaisen Teatterin lauluosasto
produced by
role in
scop:PerformanceRole
scop:Person Strömer Emmy
performer
scop:Place Arkadia-teatteri
performed in
ticket information
role in
...
additional information
"Huge success. About the performance and reception see [...]"@en
"4,00–0,50 mk"@en
performance date
start
end
xsd:date 1874-04-12
xsd:date 1874-04-12

**Figure 2.** A chart illustrating the OPERASAMPO data model through an example

performances in the scope of OPERASAMPO (opera and music theatre performances in Finland in 1830–1960) are even catalogued in the data outside of a few exceptions where the performances were not added to the database for an unknown reason. Due to this, the distinction between the original composition and the version used in the performances was not deemed necessary to be modeled in the case of the original database (Reprises) and consequently the LD version of the data follows this same simplified model as well as opposed to a FRBR-like [28] model.

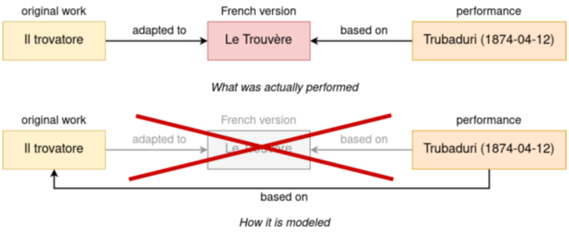


**Figure 3.** A chart illustrating the modeling decision made for performances and compositions

## 5. User Interface & Data Service

This section introduces how the data can concretely be used. The first subsection goes over the UI built on top of the data for easy searching, browsing, and analyzing of the data without the need for much technical expertise. The next subsection goes over some of the other ways the data can be utilized by using the open SPARQL endpoint for more technically knowledgeable researchers and members of the public.

**User Interface** The OperaSampo Portal[24] is built with the Sampo-UI framework[25] [2,3]. The contents of the portal are based on the wants and needs of the performance researchers: maintaining and using the original Reprises data. The goal here was to enhance its search capabilities to a new level with the inclusion of faceted search and browsing [29,30] with integrated data-analytic tools, as suggested in the Sampo model. In addition to the actual portal, a video showcasing its basic functionality is also available on the Web[26].

A Sampo-UI interface is split into different *application perspectives* for zooming into the data. The perspectives can be selected on the landing page of the portal. Each perspective utilizes the same data but presents it from a different point of view. Each perspective is based on a class of the data model and is used to search and analyze instances of the class, e.g., instances of performances.

The OperaSampo Portal has five perspectives available:

1. *Performances*. This perspective presents the data from the perspective of individual performances, e.g., *Trubaduri* (Verdi's *Il trovatore*) performed on date April 12, 1874 at Arkadia Theatre.
2. *People*. This perspective presents the data from the perspective of people involved in different aspects of the data, e.g., the performers performing in different roles or the composers of compositions.
3. *Compositions*. This perspective presents the data from the perspective of compositions, e.g., Verdi's *Il trovatore* or Wagner's *Tannhäuser*.
4. *Role characters*. This perspective presents the data from the perspective of role characters of compositions, e.g., the *Conte di Luna* from Verdi's *Il trovatore* or *Biterolf* from Wagner's *Tannhäuser*.
5. *Performance venues*. This perspective presents the data from the perspective of venues, where performances have been performed, e.g., *Svenska Teatern i Helsingfors* or *Alexander Theatre*.

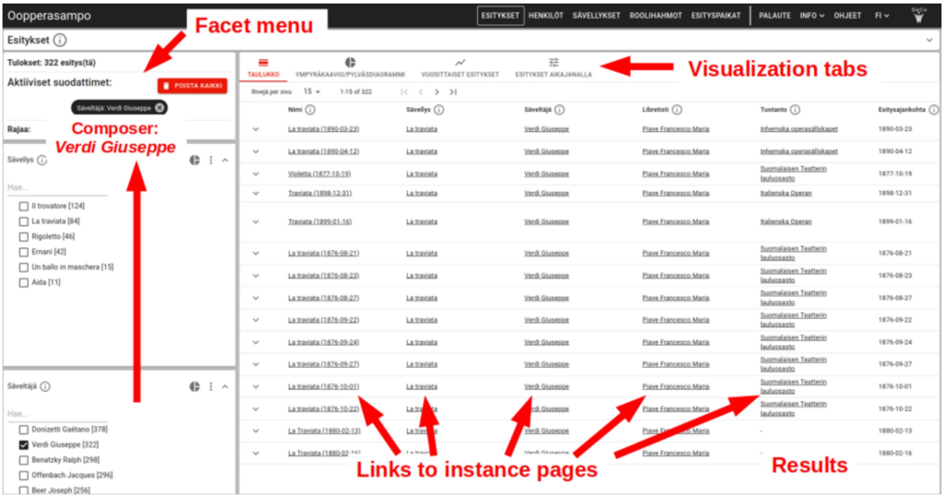
In addition to these perspectives, there are two text-content pages available on the landing page: *Sources* and *Links*. These pages provide additional information on the sources used as well as links to other archives and databases, respectively.

Selecting a perspective opens up that perspective's *faceted search view* (shown in Figure 4). The results are shown on the right side of the screen with the faceted search menus on the left and different available visualization tabs listed on top of the results. By default, the results include all entities matching the facet class of the perspective (e.g., all performances in the perspective for performances), each entity having its own row. In Figure 4, 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. The columns of the table represent the different properties and their values that the entities have. After each selection, the hit counts of all facet categories are updated showing the size of the result set if a category is selected next. The hit counts direct the search to only directions where results can be found, and can be used as a basis for statistical distribution analyses of the result set along the facets.

[24] Source code available at: https://github.com/SemanticComputing/operasampo-web-app

[25] Source code available at: https://github.com/SemanticComputing/sampo-ui

[26] https://vimeo.com/805493196



| Nimi | Sävellys | Säveltäjä | Libretisti | Tuotanto | Esitysajankohta |
|---|---|---|---|---|---|
| La traviata (1890-03-23) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Inhemska operasällskapet | 1890-03-23 |
| La traviata (1890-04-12) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Inhemska operasällskapet | 1890-04-12 |
| Violetta (1877-10-19) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1877-10-19 |
| Traviata (1898-12-31) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Italienska Operan | 1898-12-31 |
| Traviata (1899-01-16) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Italienska Operan | 1899-01-16 |
| La traviata (1876-08-21) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-08-21 |
| La traviata (1876-08-23) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-08-23 |
| La traviata (1876-08-27) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-08-27 |
| La traviata (1876-09-22) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-09-22 |
| La traviata (1876-09-24) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-09-24 |
| La traviata (1876-09-27) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-09-27 |
| La traviata (1876-10-01) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-10-01 |
| La traviata (1876-10-22) | La traviata | Verdi Giuseppe | Piave Francesco Maria | Suomalaisen Teatterin lauluosasto | 1876-10-22 |
| La Traviata (1880-02-13) | La traviata | Verdi Giuseppe | Piave Francesco Maria | - | 1880-02-13 |
| La Traviata (1880-02-16) | La traviata | Verdi Giuseppe | Piave Francesco Maria | - | 1880-02-16 |

**Figure 4.** The faceted search view of the Performances perspective

Underlined text in the results table indicates a link to either internal or external (e.g., Wikidata) pages. For most of the cases, these links lead to *instance pages* of entities, e.g., a page for a particular performance. These pages list aggregated linked information available about that particular entity and might include various visualizations for it if they are available for entities of that particular type.

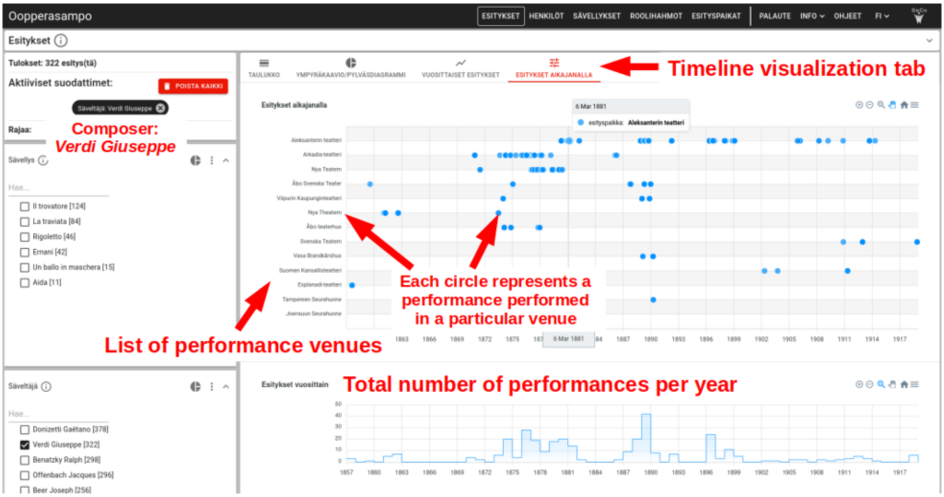
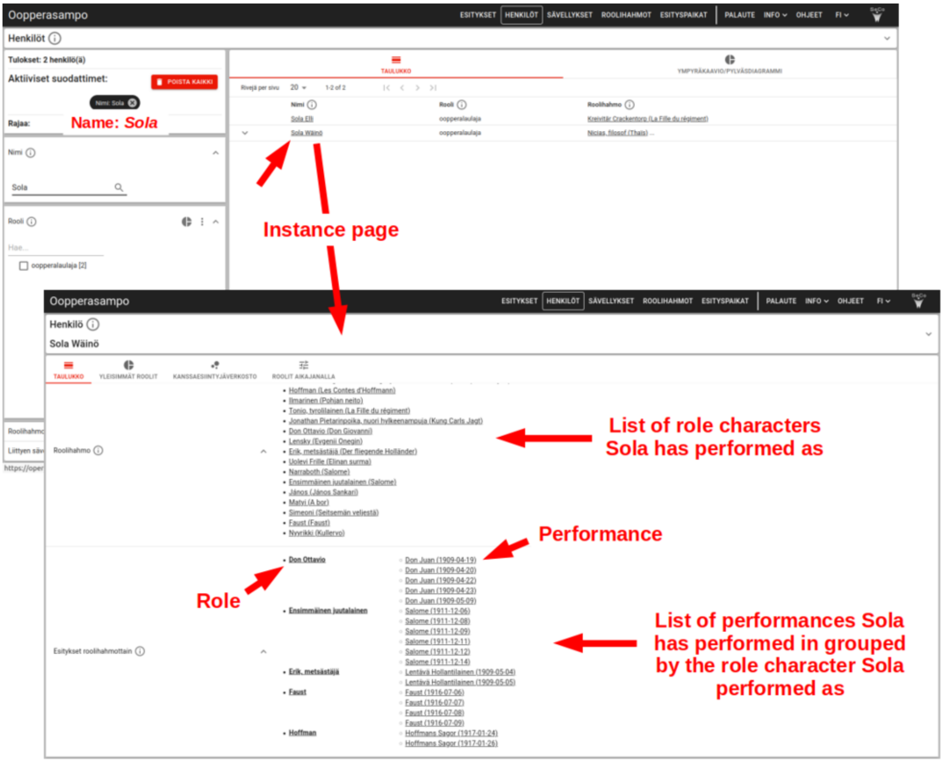


**Figure 5.** A visualization illustrating the annual number of performances as well as the venues they were performed in

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 with 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. This kind of detailed information is valuable in researching historical performance culture, and through the portal it is now easy to access.



Tulokset: 2 henkilö(ä)

Aktiiviset suodattimet:

Rajaa:

Nimi: Sola

Rooli: Hae... □ oopperalaulaja [2]

Rivejä per sivu 20 1-2 of 2

| Nimi | Rooli | Roolihahmo |
|---|---|---|
| Sola Elli | oopperalaulaja | Kreivitär Crackentorp (La Fille du régiment) |
| Sola Wäinö | oopperalaulaja | Nicias, filosof (Thaïs) ... |

Sola Wäinö

Roolihahmo

- Hoffman (Les Contes d'Hoffmann)
- Ilmarinen (Pohjan neito)
- Tonio, tyrolilainen (La Fille du régiment)
- Jonathan Pietarinpoika, nuori hylkeenampuja (Kung Carls Jagt)
- Don Ottavio (Don Giovanni)
- Lensky (Evgenii Onegin)
- Erik, metsästäjä (Der fliegende Holländer)
- Uolevi Frille (Elinan surma)
- Narraboth (Salome)
- Ensimmäinen juutalainen (Salome)
- János (János Sankari)
- Matyi (A bor)
- Simeoni (Seitsemän veljestä)
- Faust (Faust)
- Nyyrikki (Kullervo)

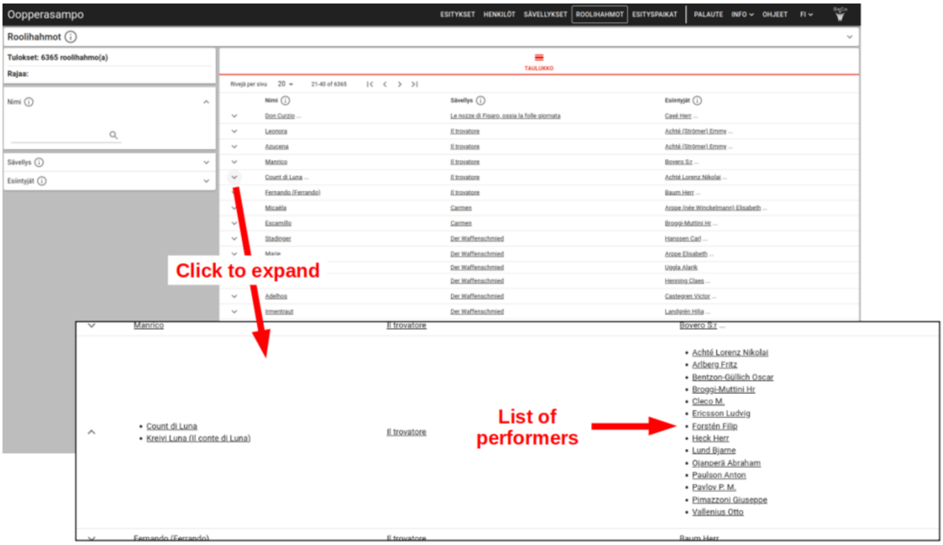
Esitykset roolihahmottain

- Don Ottavio
  - Don Juan (1909-04-19)
  - Don Juan (1909-04-20)
  - Don Juan (1909-04-22)
  - Don Juan (1909-04-23)
  - Don Juan (1909-05-09)
- Ensimmäinen juutalainen
  - Salome (1911-12-06)
  - Salome (1911-12-08)
  - Salome (1911-12-09)
  - Salome (1911-12-11)
  - Salome (1911-12-12)
  - Salome (1911-12-14)
- Erik, metsästäjä
  - Lentävä Hollantilainen (1909-05-04)
  - Lentävä Hollantilainen (1909-05-05)
- Faust
  - Faust (1916-07-06)
  - Faust (1916-07-07)
  - Faust (1916-07-08)
  - Faust (1916-07-09)
- Hoffman
  - Hoffmans Sagor (1917-01-24)
  - Hoffmans Sagor (1917-01-26)

**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

Other particularly interesting part of the data from the perspective of the DH researchers 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, 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 a certain role char-



Tulokset: 6365 roolihahmo(a)

Rajaa:

| Nimi | Sävellys | Esiintyjät |
|---|---|---|
| Don Curzio ... | Le nozze di Figaro, ossia la folle giornata | Cavé Herr ... |
| Leonora | Il trovatore | Achté (Strömer) Emmy ... |
| Azucena | Il trovatore | Achté (Strömer) Emmy ... |
| Manrico | Il trovatore | Bovero S:r ... |
| Count di Luna ... | Il trovatore | Achté Lorenz Nikolai ... |
| Fernando (Ferrando) | Il trovatore | Baum Herr ... |
| Micaëla | Carmen | Arppe (née Winckelmann) Elisabeth ... |
| Escamillo | Carmen | Broggi-Muttini Hr ... |
| Stadinger | Der Waffenschmied | Hanssen Carl ... |
| Marie | Der Waffenschmied | Arppe Elisabeth ... |
|  | Der Waffenschmied | Uggla Alarik |
|  | Der Waffenschmied | Henning Claes ... |
| Adelhos | Der Waffenschmied | Castegren Victor ... |
| Irmentraut | Der Waffenschmied | Landgrén Hilja ... |

| | | |
|---|---|---|
| Manrico | Il trovatore | Bovero S:r ... |
| Count di Luna<br>Kreivi Luna (Il conte di Luna) | Il trovatore | Achté Lorenz Nikolai<br>Arlberg Fritz<br>Bentzon-Güllich Oscar<br>Broggi-Muttini Hr<br>Cleco M.<br>Ericsson Ludvig<br>Forstén Filip<br>Heck Herr<br>Lund Bjarne<br>Ojanperä Abraham<br>Paulson Anton<br>Pavlov P. M.<br>Pimazzoni Giuseppe<br>Vallenius Otto |
| Fernando (Ferrando) | Il trovatore | Baum Herr |

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

acter. 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*).

**Data Service** The OperaSampo dataset[27] is available on the Linked Data Finland[28] platform with a CC BY 4.0 license. The triples are stored on a Apache Jena Fuseki[29] SPARQL server that is accessible from an open SPARQL endpoint[30]. The data uses Lucene[31] text indexing. Using the open SPARQL endpoint, the user can query all the data included in the OperaSampo KG and visualize it using, e.g., SPARQL queries on the Yasgui[32] SPARQL client [31] or with Python using the Google Colab environment[33]. This enables the user to create more complex queries on the data that might not be possible to create with the UI.

Figure 8 illustrates an example of using the Yasgui SPARQL client for visualizing the data outside of the UI. In this particular example, the user has created a query that fetches all the performances for compositions that have been composed by the composer *Emmerich Kálmán*[34], groups them by the decade based on performance date, and counts the number of performances. This information has been visualized using the Yasgui SPARQL client to depict the information as a stacked bar chart with decades making up the X-axis and the number of performances the Y-axis. Different compositions, the different series in the graph, are highlighted with different colors. Alternatively, the

[27] https://www.ldf.fi/dataset/operasampo
[28] https://www.ldf.fi/
[29] https://jena.apache.org/documentation/fuseki2/
[30] https://ldf.fi/operasampo/sparql
[31] https://lucene.apache.org/
[32] http://yasgui.triply.cc/
[33] https://colab.research.google.com/
[34] This is the composer whose compositions have the highest total count of performances performed in the OperaSampo KG.

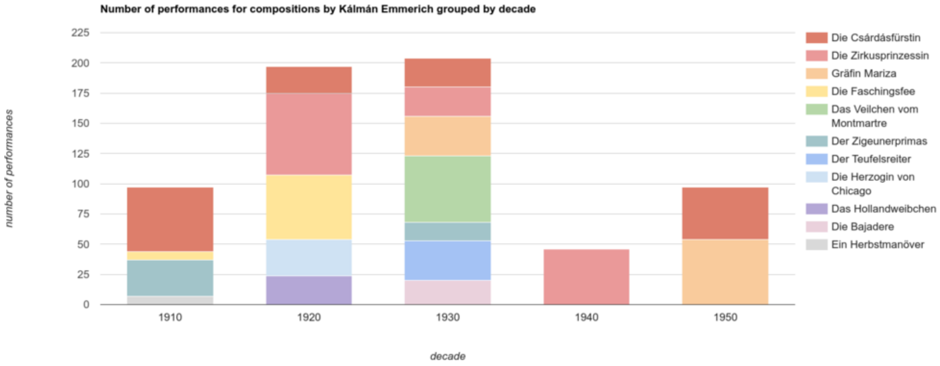


**Figure 8.** A visualization of number of performances for a composition composed by a particular composer created using Yasgui SPARQL client.

user could have downloaded the query results and visualized them by, e.g., using Python scripts and libraries.

In performance research this kind of visualization is a very useful tool for grasping quickly an overview of the reception process: the frequency of performances of the pieces (here operettas) and their chronology: which pieces were performed the earliest, which came after, which were performed most frequently, which during a particular time period. But, for a full view of the Finnish Kálmán reception, also Encore, The Finnish National Opera and Ballet performance database needs to be consulted.

## 6. Data Maintenance

The OperaSampo data is maintained in LD format with the SAHA metadata editor [32]. Some initial data corrections, such as missing performance names, were initially either corrected using the original Reprises database editing interface. They were then brought over to the created LD data repository by automatically converting newer data dump versions to the correct format using the same Python scripts that were used to create the initial LD version of the data. Later on some corrections were also included in the conversion steps themselves, when correcting things was no longer feasible on the original Reprises database. After the SAHA metadata editing interface was set up, the data corrections were done to the LD format data exclusively.

Some of the issues in the data were made more glaring through the new user interface and how it shows information. Both the table view of the data as well as the facets for different properties in the user interface make missing data or scarcely annotated properties more apparent, as this information was previously hidden behind the 'cards' for specific performances and could not be used for searching in the same capability as with the new user interface. Some apparent issues were selected as the priority to fix before the portal was officially made open to all users: (1) information on performance venues, and (2) duplicate instances of the same person.

There's a limited amount of performance venues in the data (around 60) and the *Performance venues* perspective was assumed to be one of the perspectives where people might look at the data more closely, so this data was heavily supplemented with additional information (new/more exact venues as well as additional information to the existing venues, e.g., addresses, other names, short description texts) and pictures as the original data was very bare-bones. New and/or more exact venues were added to replace previous more ambiguous ones, e.g., an ambiguous 'Tampere' (a city in Finland) instance was replaced with the information on the actual venue located in Tampere where the performances that were linked to it were actually performed. For existing venues new pictures from throughout their histories were added as well as information on the addresses of venues, if they were known. Short text descriptions were also added for venues for the purpose of enlightening possible users on the history of these venues.

During this performance venue information supplementation process it was also decided to merge some venue entities together. In the original Reprises data venues that went through name changes were split into multiple entities. For example, the venue nowadays know as *Svenska Teatern i Helsingfors*[35] was previously known as *Nya Theatern* and *Nya Teatern*. The performances performed in this venue were thus split between these three entities instead of being aggregated under a single entity. Even though the building itself has changed from its original form, the theater is still located in the same place and can be thought of as a single theater with a long, rich history. Cases like these were combined under a single venue entity with the previous names being listed as alternative labels for the venues. This way the user can search for all performances performed in a venue without having to know the exact history of a venue. For example, the user can filter out all performances performed in Svenska Teatern by selecting just that in, e.g., a facet, without having to be aware enough of the history of the venue to also select Nya Theatern and Nya Teatern. And should this information be relevant, it can be enhanced by consulting the performance venue perspective where short histories are provided for.

In a similar vein to the combined performance venues, evident duplicate person entities were also combined together. There were some cases in the data where some people mistakenly had two entities with the exact same name and information (e.g., year of birth). These cases were just combined into one without any changes to the information. Another type of case was people that had changed their name at some point in their life or had a stage name they used. The focus on these was on the women, who were more likely to have changed their surnames during their career. Information on the people with the same first names was queried from the database and compared together to find possible duplicates. Many of the duplicate cases had one of the entities mentioning a future or a previous surname, so they could be flagged based on that for further investigation. Some other cases flagged for further investigation weren't as obvious and were based on when the people had been shown to be active and/or alive, and in some cases, having similar enough names that one of them could just be a typo or an alternative spelling. These flagged cases were looked into and combined if necessary based on the findings. For these cases the alternate names were also included in the new combined person instance as alternative or hidden labels depending on the nature of the alternate name (e.g., previous name vs. a rare alternative spelling).

New, previously uncatalogued performances will be added as time goes on. New data is being added by people who were maintaining the original database after SAHA

[35] The theater has been known as *Svenska Teatern* (name in Swedish) or *Ruotsalainen teatteri* (name in Finnish) since 1888. The original name of the theater inaugurated in 1860 was *Nya Theatern*, before it was changed to *Nya Teatern* in 1870 and later on to *Svenska Teatern*.

learning sessions were hosted for these people. New people are also trained in the usage of SAHA for possible larger, new additions to the data, such as music theatre performances of the Finnish Operetta Theater, touring performance groups, Sibelius Academy Opera Studio, as well as those of the Jewish Dramatic Society in Finland, to mention a few. Data enriching links to other sources, such as Wikidata and BiographySampo [33] have been and continue to be added for the people in the KG to make it possible use data from other sources.

## 7. Evaluation

An evaluation of the UI was carried out to assess its usability by one its user demographics, researchers working with historical opera and music theatre performance data. As the evaluators were not necessarily familiar with the portal before, basic usage instructions were sent to all the possible evaluators with the link to the evaluation survey.

The evaluation consisted of the end user first having to answer some relatively simple questions using the UI, e.g., who is the performer with the most performances in a specific role or what is the composition with most performances in the data. The second part of the evaluation consisted of questions to evaluate the user experience—ease of use, clarity of instructions, usefulness etc.— during the process of looking for the answers to the previous questions. This section also allowed the users to leave free-form feedback regarding the portal.

The first few easy tasks that had the user explore the *Performances* perspective were correctly answered by all the respondents, but there was some variance in the answers given to the more complex tasks that required the user to either manipulate the information order (e.g., sorting based on date) or navigate the other perspectives and instance pages to find the information.

The number of correct answers to the first section of the evaluation seemingly correlated with the respondent's satisfaction with the general use experience as well as the instructions included; the people that were able to answer all the questions correctly rated the clarity of the instructions higher (7.5 average on the scale 0–10) while those that struggled with the more complex tasks rated the clarity lower (4.5 average). In the free-form feedback, there were a few mentions of needing some more time with the UI for it to potentially start feeling intuitive to use. Navigation between pages and visualization-specific controls were specifically mentioned as difficult by some of the respondents.

Overall, the portal was found to be potentially useful for the end users in their work (8.0 average) and received good overall ratings for the whole experience (8.3 average). An empirical indication for its usefulness and usability is the fact that it has already been used to aid research in the field of Finnish historical opera and music theatre performances, too. The overall evaluation is also in line with the evaluation done to another Sampo-UI-based portal. This other portal, the Mapping Manuscript Migrations (MMM) portal [34], was evaluated in [35] and suggested promising usability of the underlying Sampo model as well as the UI logic from the perspective of an end user. The usage of Sampo systems in various CH portals[36] with some having more than a million annual users is also an empirical indication of the usability of the model.

[36] Information about the Sampo portal series is available at: https://seco.cs.aalto.fi/applications/sampo/

From a computation standpoint, the Sampo-UI is scalable enough to handle up to hundreds of thousands of instances, though this is affected by the complexity of the underlying data model [2]. The Sampo-UI also enables the developer to force the user to constrain the data set before performing computationally heavy tasks like the hit count computation for all the values in different facet categories that must be updated after each new facet selection. This forcing of constraints is used in, for example, NameSampo [36] to handle the data set with over two million placenames.

## 8. Discussion

In contrast to the related works, the novelty in OPERASAMPO lays on its focus on using historical data to study the musical performances and persons involved in different roles instead of focusing on the composers, compositions, and premieres. Instead of performers just being strings of names listed on a performance's page like in many traditional opera databases (or even less), each performer is its own instance aggregating information about that particular performer.

The OPERASAMPO PORTAL offers a novel way of exploring and searching the performance data in a way that was not possible before with the *Reprises* database's text-based search and limited filtering options. In comparison to traditional text-based search, 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. The linked nature of the data makes it easier for the user to jump from different instances to another, e.g., from a performance to the page of a specific performer. In addition, the data is enriched by data linking, the scope of which is being gradually expanded during the continued data maintenance and addition of new data to the KG.

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. The visualizations update with the results set, so the user can easily, for example, compare different subsets of data (e.g., performances performed in different performance venues) without having to necessarily extract the query data to be used with, e.g., Python scripts. For those looking to formulate more complex queries and data analyses, the open SPARQL endpoint can be used for querying the necessary data.

The evaluation suggested a promising usability of the portal for researchers working in the field of historical opera and music theatre performances, with some room to improve on the instructions for the portal and its usage. As the portal was developed in direct collaboration with researchers already utilizing the underlying data, the development has been guided by the needs of this demographic. The portal has already been used as an aid during research on historical performances to quickly and easily get an overview of when and how many times a particular composition or compositions have been performed without having to manually find and count them from, e.g., a physical repertoire book or newspaper advertisements for performances, and being able to further filter these results based on, e.g., a certain performer being present in them.

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

## References

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

[2] Ikkala E, Hyvönen E, Rantala H, Koho M. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability. 2022;13(1):69-84.

[3] Rantala H, Ahola A, Ikkala E, Hyvönen E. How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In: Proceedings of 8th International Workshop on the Visualization and Interaction for Ontologies and Linked Data co-located with the 22nd International Semantic Web Conference (ISWC 2023) in Athens, Greece; 2013. Accepted, forth-coming.

[4] Hyvönen E. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web – Interoperability, Usability, Applicability. 2023;14(4):729-44.

[5] Ahola A, Hyvönen E, Rantala H, Kauppala A. Publishing and studying historical opera and music theatre performances on the Semantic Web: case OperaSampo 1830–1960. In: Proceedings of SWODCH 2023. Semantic Web and Ontology Design for Cultural Heritage. Co-located with the 22nd International Semantic Web Conference (ISWC 2023) in Athens, Greece. CEUR Workshop Proceedings, Vol-3540; 2023. Available from: https://ceur-ws.org/Vol-3540/paper8.pdf.

[6] Hyvönen E, Ahola A, Rantala H, Kauppala A. OperaSampo – Opera and Music Theatre Performances in Finland 1830–1960 on the Semantic Web. In: Fundulaki I, Kozaki K, Gomez-Perez JM, Garijo D, editors. Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 21st International Semantic Web Conference (ISWC 2023). CEUR Workshop Proceedings; 2023. .

[7] Tietz T, Waitelonis J, Zhou K, Felgentreff P, Meyer N, Weber A, et al. Linked Stage Graph. In: SEMANTICS Posters&Demos. vol. 2451. CEUR Workshop Proceedings, http://CEUR-WS.org; 2019. Available from: https://ceur-ws.org/Vol-2451/paper-27.pdf.

[8] Lesnig G. Die Aufführungen der Opern von Richard Strauss im 20. Jahrhundert: Daten, Inszenierungen, Besetzungen, Band 2. Hans Schneider, Tutzing; 2010.

[9] Pattuelli MC, Hwang K, Miller M. Accidental discovery, intentional inquiry: Leveraging linked data to uncover the women of jazz. Digital Scholarship in the Humanities. 2016 10;32(4):918-24. Available from: https://doi.org/10.1093/llc/fqw047.

[10] Morales Tirado A, Carvalho J, Ratta M, Uwasomba C, Mulholland P, Barlow H, et al. Musical Meetups Knowledge Graph (MMKG): a collection of evidence for historical social network analysis. In: European Semantic Web Conference. Springer; 2024. p. 110-27.

[11] de Berardinis J, Carriero VA, Jain N, Lazzari N, Meroño-Peñuela A, Poltronieri A, et al. The polifonia ontology network: Building a semantic backbone for musical heritage. In: International Semantic Web Conference. Springer; 2023. p. 302-22.

[12] Achichi M, Lisena P, Todorov K, Troncy R, Delahousse J. DOREMUS: A graph of linked musical works. In: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17. Springer; 2018. p. 3-19.

[13] Bechhofer S, Page K, De Roure D. Hello cleveland! Linked data publication of live music archives. In: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). IEEE; 2013. p. 1-4.

[14] Bechhofer S, Page KR, Weigl DM, Fazekas G, Wilmering T. Linked Data Publication of Live Music Archives and Analyses. In: The Semantic Web – ISWC 2017; 2017. p. 29-37.

[15] Page KR, Bechhofer S, Fazekas G, Weigl DM, Wilmering T. 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. p. 1-10.

[16] Devaney J, Gauvin HL. Representing and Linking Music Performance Data with Score Information. In: Proceedings of the 3rd International Workshop on Digital Libraries for Musicology. DLfM 2016. New York, NY, USA: Association for Computing Machinery; 2016. p. 1–8. Available from: https://doi.org/10.1145/2970044.2970052.

[17] Bonora P, Pompilio A. RePIM in LOD: semantic technologies to manage, preserve, and disseminate knowledge about Italian secular music and lyric poetry from the 16th-17th centuries. Umanistica Digitale. 2022;(14):71-90.

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

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

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

[21] van Nieuwkerk MM, Salters L, Helmers RM, Kisjes I. Operatic Productions in the Netherlands, 1886–1995: from Printed Annals to Searchable Performance Data. Research Data Journal for the Humanities and Social Sciences. 2020;5:79-90. Available from: https://brill.com/view/journals/rdj/5/2/article-p79_79.xml?language=en.

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

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

[24] Hitzler P, Krötzsch M, Rudolph S. Foundations of Semantic Web technologies. CRC press; 2009.

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

[26] Gademan G. Realismen på operan. Regi, spelstil och iscensättningsprinciper på Kungliga Teatern 1860–62. Stockholm: Teatervetenskapliga institutionen; 1996.

[27] Paavolainen P. Nuori Bergbom: Kaarlo Bergbomin elämä ja työ. I, 1843-1872; 2014. Available from: https://urn.fi/URN:ISBN:978-952-6670-23-2.

[28] Tillett B. What is FRBR? A conceptual model for the bibliographic universe. The Australian Library Journal. 2005;54(1):24-30.

[29] Hearst M. Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR workshop on faceted search. Seattle, WA; 2006. p. 1-5.

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

[31] Rietveld L, Hoekstra R. The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability. 2017;8(3):373-83.

[32] Mäkelä E, Hyvönen E. SPARQL SAHA, a Configurable Linked Data Editor and Browser as a Service. In: Proceedings of the ESWC 2014 demonstration track, Springer-Verlag; 2014. .

[33] Hyvönen E, Leskinen P, Tamper M, Rantala H, Ikkala E, Tuominen J, et al. 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. p. 574-89.

[34] Hyvönen E, Ikkala E, Koho M, Tuominen J, Burrows T, Ransom L, et al. Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research. In: Semantic Web. Proceedings of the The 20th International Semantic Web Conference (ISWC 2021). Springer; 2021. .

[35] Burrows T, Pinto NB, Cazals M, Gaudin A, Wijsman H. Evaluating a semantic portal for the "Mapping Manuscript Migrations" project. DigItalia. 2020;15(2):178-85.

[36] Ikkala E, Tuominen J, Raunamaa J, Aalto T, Ainiala T, Uusitalo H, et al. NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research. In: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities. GeoHumanities'18. New York, NY, USA: ACM; 2018. p. 2:1-2:9.