Towards Linked Stage Graph 2.0 - a Knowledge Graph based Research Resource for the Performing Arts

Tabea Tietz[1,2], Harald Sack[1,2]
[1]FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Germany
[2]KIT Karlsruhe (AIFB), Germany
[1]{firstname.lastname}@fiz-karlsruhe.de
[2]{firstname.lastname}@kit.edu

1. Introduction

Theatre as an art form and as a social institution has been deeply embedded into our societies. Theatre is a realm that effortlessly merges the historic and the incredibly modern, continuously evolving and reinventing itself time and time again. Historical plays, which persist to this day, remain remarkably relevant, finding new life through reinterpretation on modern stages. Theatre never exists in isolation and is always embedded in the context of the respective culture and society. These contexts can be shaped by societal and political events or by technological advancements and these contexts also have shown to influence the creative possibilities and limitations on stage. Scholars from various disciplines have been examining the history and development of theatre in all its facets, including stage design, characters, costumes, and language.

A number of tools have been created with the intention to provide access to performing arts data collections. Linked Stage Graph (LSG) as one of these efforts enables the exploration of photographs and metadata of the Stuttgart State Theatre's from the 1880s until the 1940s (Tietz, 2019). The knowledge graph (KG) is accessible on the Web by means of a public SPARQL Endpoint[1] and an exploration interface[2] . While the cultural heritage community has been receiving LSG generally well, there have also been challenges which still have to be resolved for the resource to be even more useful for the cultural heritage community. For instance, the underlying ontology requires an update to provide more meaningful relations and to improve interoperability. This includes: A more accurate distinction between the original work and the performance event and an improvement of the representation of individual roles and functions of persons active. Furthermore, Linked Stage Graph currently contains textual descriptions of the archival objects. This is difficult to integrate into exploration environments in a useful and meaningful way. An improved extraction of notable entities and their integration into the KG enables more efficient querying.

This contribution reports on the ongoing process of creating a new version of LSG as a technically improved and even more useful research resource for the culture community. Goal of this paper is to provide a strategy towards Linked Stage Graph 2.0 (Section 4.2) based on systematically extracted requirements (Section 3) and by taking into account the challenges (Section 4.1) of the data set. These contributions are not only limited to LSG and concern performing arts data in general. Therefore, this

[1] https://slod.fiz-karlsruhe.de/sparql
[2] https://slod.fiz-karlsruhe.de/vikus

paper is furthermore an invitation to discuss current systems and data models as well as strategies to more forward in the performing arts sector from a more technical point of view.

2. Related Work

There are a number of systems, platforms and data models which provide access to (historical) performing arts data. Often, their entry point is focused on either a biographical, a regional or an institutional approach. LSG is based on historical data created at Stuttgart State Theatres and provided by the Baden-Württemberg National Archives.

Biographical approaches include Ipsen Stage[3] , the Pina Bausch Archive (Thull, Diwisch and Marz, 2015) and Staging Beckett (McMullan et al., 2014). Furthermore, Re-Collecting Theatre History (Probst and Pinto, 2020) provides a data capturing tool and an exploration environment for theatre sciences. Approaches with a more institutional focus include the Abby Theater Platform (Bradley and Keane, 2015) and the Specialised Information Service for Performing Arts (Beck et al., 2017). Important regional approaches include the Dutch project ONSTAGE (Blom, Nijboer and Zalm, 2020), AusStage (Bollen, 2016), the Swiss Performing Arts Platform (Estermann and Julien, 2019) and OperaSampo (Ahola et al., 2023).

The above listed projects and platforms are valuable and relevant research resources in the performing arts domain. However, to the best of our current knowledge, there are no public SPARQL endpoints available for any of them. Even in cases where ontologies were designed to represent the data, they were not made publicly available for reuse. The SPA project is relevant for LSG, as it provides a comprehensive and well-documented published data model. Linked Stage Graph is accessible by means of a SPARQL endpoint and all data and development progress is open and documented on GitHub.[4]

3. Requirements

A requirement analysis has been carried out to improve LSG systematically. The requirements were extracted from scientific literature in the domain of the performing arts. All requirements can be traced down to the literature they were extracted from and the full overview is available on GitHub. This analysis is an ongoing process and will be extended as part of iterative ontology development.

REQ1: Context. Provide as much context as possible. No entity in the performing arts exists on its own (e.g. archival object, performance, person, prop) and should be viewed in its context to other entities.

REQ2: Perspective. Provide various perspectives on performing arts data for exploration to enable a holistic view.

REQ3: Interoperability. Enable the interconnection between disciplines, data sets, archives, performing arts institutes as well as regional and international efforts.

REQ4: Persons. All persons on, behind and in front of the stage of a performance and their roles and functions are relevant for research.

REQ5: Change. Performing arts are dynamic and the change over time should be represented in terms of persons, occupations, stage design, etc.

[3] https://ibsenstage.hf.uio.no/
[4] https://github.com/ISE-FIZKarlsruhe/LinkedStageGraph

REQ6: Events. Performing arts data is often event-based. It should be distinguished between an original work, a production and the performance as an event.

REQ7: Stage Elements. Objects on stage should be captured. If possible, the meaning of an object on stage should be represented as well.

REQ8: Querying. A data model that represents performing arts data should be as lightweight as possible to enable intuitive querying.

REQ9: Provenance. It has to be possible to verify and track research results, e.g. biographical data has to be linked to their data source.

REQ10: Data Quality. The quality of the data used in performing arts research has to be clear and should be quantifiable.

These requirements are being carefully taken into account throughout the development process. However, not all requirements can be fully met due to the sparsity of the available metadata.

4. Towards Linked Stage Graph 2.0

This chapter discusses additional challenges within LSG and provides a strategy with concrete tools and standards to overcome them and meet the requirements.

4.1. Current Challenges

In the following, the most relevant challenges within the LSG data are described.

CH1: Archival Structure. The original data structure is based on the folders in the archive before digitization and not intuitive for web-based exploration. These folders reveal information about the carrier type of the photographs in the data set (e.g. glass plate), but also the performance types (e.g. ballet).

CH2: Semi-structured Data. The titles and descriptions of the archival objects contain many unstructured information about the performances, dates, and persons involved.

CH3: Performance Categories. Performances are categorized by genre (e.g. opera, ballet) and by type (e.g. premiere, repertoire) in a semi-structured way. Between these concepts relations exist (opera, romantic opera) but they are not reflected in the original data set. So far, no reference lists could be determined for a unique identification.

CH4: Heterogeneity. Not every archival object in the data set is part of a performance, e.g. outdoor photographs of theatre buildings.

CH5: Metadata Sparsity. Performance data often lack metadata, e.g. actors are often not listed and performance dates are sometimes missing.

Even though these challenges were observed within LSG, they are not limited to this data set only and are generalizable to historical German archival data.

4.2. Strategy

According to the requirements, performing arts research data should be interconnected, and interoperable with as much context provided as possible, and should represent events and change meaningfully. To

fulfil these needs, KGs present a state-of-the-art solution. The following sections discuss the intended strategy in order to fulfil the listed requirements as part of a KG-based solution.

4.2.1. Data Model

To the best of our knowledge, the Swiss Performing Arts (SPA) data model is the most semantically expressive and best documented model that is available in the performing arts domain (Estermann and Julien, 2019). It is event based and emphasizes the importance of differentiating the original work and the performance event (REQ6) by reusing CIDOC, FRBR and FRBRoo. SPA uses RiC-O to represent the archival structure, which will be utilized to cope with CH1 and REQ9. However, the SPA is incredibly complex (REQ8) and assumes rich metadata, which is not present within LSG. For instance, there is no property between the work[5] and the performance event[6] itself and all persons and their functions (e.g stage designers) are connected with the production. However, in Linked Stage Graph it is not clear which production a performance event belongs to. Therefore, all persons are linked to the respective performance event. The model further connects the Performing Arts Production with a Performance Plan[7] and Performance Work[8] which are not available in the Linked Stage Graph metadata. Therefore, the model cannot be reused entirely, but will partially be utilized. Modeling challenges and strategies are furthermore discussed in Tietz et al. (2023).

4.2.2. Semi-structured Data

To cope with CH2, Named Entity Recognition and Linking will be conducted to extract mentions of persons and their functions, works, and dates from semi-structured descriptions of the archival objects. A mapping with existing data sources like GND an Wikidata will improve interoperability (REQ3) and enables to provide more context to the data by means of enrichment (REQ1). Furthermore, NIF2 (Hellmann et al., 2013) can be used to provide confidence levels (REQ10).

4.2.3. Perspective

As mentioned in CH5 for many archival objects within Linked Stage Graph no rich metadata are available. However, to meet REQ2 and provide a holistic data exploration environment, sufficient metadata is crucial. Therefore, the photographs within the collections are analyzed by means of state-of-the-art object detection and caption generation (REQ7). Previous experiments (Tietz et al., 2020) have revealed the challenges within this task. To integrate the image analysis results into the KG the Web Annotation Ontology[9] can be applied. As a consequence, means of explorations can exceed the current timeline-based approach and also allow searching within the photographs.

4.2.4. Categorizing Performances

The categorization of performances by type and genre as mentioned in CH3 is crucial for data exploration. However, to the best of our knowledge no reference lists exist which are widely accepted by the performing arts community to relate individual concepts (e.g. opera, comic opera, comedy, romantic

[5] SPA-E2: https://www.iflastandards.info/fr/frbr/frbroo#F1
[6] SPA-E9/10: https://www.iflastandards.info/fr/frbr/frbroo#F31
[7] SPA-E8: https://www.iflastandards.info/fr/frbr/frbroo#F25
[8] SPA-E7: https://www.iflastandards.info/fr/frbr/frbroo#F20
[9] https://www.w3.org/TR/annotation-vocab/

opera) to each other. However, the Art and Architecture Thesaurus[10] can be reused partially for this task. Furthermore, a mapping with Wikidata entities is utilized.

4.2.5. Context and Interoperability

Publishing performing arts data by means of a KG and connecting individual entities, e.g. persons, organizations, events to existing sources like the GND or Wikidata also provides the opportunity to enrich the existing data set with further context (REQ1). As mentioned, the performing arts are also embedded into societal changes and more context enables us to answer questions like: Which theatres performed plays by Bertolt Brecht during the 1930s?. Furthermore, the KG based approach allows to easily connect the resources with further research data, e.g. via NFDI4Culture. Having registered LSG in the Culture Information Portal[11] it is possible to find useful connections to other data sets by means of federation.

This chapter outlined current challenges and means to progress and create a more scientifically and technically correct, interconnected, and open resource by reusing open standards, allowing connections with further performing arts related resources and thus, increasing its value.

5. Conclusion

This paper contributes a road map towards Linked Stage Graph 2.0 as a KG-based performing arts resource, including a requirement analysis as well as a discussion of ongoing challenges and a future strategy. This road map is not intended to be the sole truth towards an open and interoperable resource in the performing arts domain. Instead, it serves as a means to progress in the field by leveraging state-of-the-art technologies and standards. This contribution is furthermore an invitation to discuss use cases, data models and exploration environments with the cultural heritage community. The development process of Linked Stage Graph is transparent, with ongoing reporting of lessons learned. In addition, all data, ontologies, requirements created along the way are being published on GitHub.

Appendix A

Bibliography

1. **Ahola, Annastiina, Eero Hyvönen, Heikki Rantala, and Anne Kauppala.** 2023. *Publishing and Studying Historical Opera and Music Theatre Performances on the Semantic Web: Case OperaSampo 1830-1960.*" In Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH), co-located with ISWC 2023. CEUR WS Vol-3540.
2. **Beck, Julia, Michael Büchner, Stephan Bartholmei, and Marko Knepper**. 2017. *Performing Entity Facts: The Specialised Information Service Performing Arts*. Datenbank Spektrum 17 (1): 47–52.

[10] https://www.getty.edu/research/tools/vocabularies/aat/
[11] https://nfdi4culture.de/resource/E3590/about.html

3. **Blom, Frans R. E., Harm Nijboer, and Rob van der Zalm**. 2020. *ONSTAGE, the Online Data System of Theatre in Amsterdam from the Golden Age to Today*. Research Data Journal for the Humanities and Social Sciences 5 (2): 27–40.
4. **Bollen, Jonathan**. 2016. *Data Models for Theatre Research: People, Places, and Performance*. Theatre Journal, 615–632.
5. **Bradley, Martin, and Aisling Keane**. 2015. *The Abbey Theatre Digitization Project in NUI Galway.* New Review of Information Networking 20 (1-2): 35–47.
6. **Estermann, Beat, and Frédéric Julien**. 2019. *A Linked Digital Future for the Performing Arts: Leveraging Synergies along the Value Chain*. In Canadian Arts Presenting Association (CAPACOA) in cooperation with the Bern University of Applied Sciences.
7. **Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer**. 2013. *Integrating NLP using linked data*. In 12th International Semantic Web Conference, Sydney, NSW, Australia, Proceedings, Part II 12, 98–113. Springer.
8. **McMullan, Anna, Trish McTighe, David Pattie, and David Tucker**. 2014. *Staging Beckett: constructing histories of performance*. Journal of Beckett Studies 23 (1): 11–33.
9. **Probst, Nora, and Vito Pinto**. 2020. *Re-Collecting Theatre History, Theaterhistoriografische Nachlassforschung mit Verfahren der Digital Humanities*. In Neue Methoden der Theaterwissenschaft, edited by Benjamin Wihstutz and Benjamin Hoesch, 157–179. transcript.
10. **Thull, Bernhard, Kerstin Diwisch, and Vera Marz**. 2015. Linked Data im digitalen Tanzarchiv der Pina Bausch Foundation. In Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften, 259–275.
11. **Tietz, Tabea, Oleksandra Bruns, and Harald Sack.** 2023. *A Data Model for Linked Stage Graph and the Historical Performing Arts Domain*. In Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH), co-located with ISWC 2023. CEUR WS Vol-3540.
12. **Tietz, Tabea, Jörg Waitelonis, Mehwish Alam, and Harald Sack**. 2020. *Knowledge Graph based Analysis and Exploration of Historical Theatre Photographs*. In Qurator 2020.
13. **Tietz, Tabea, Jörg Waitelonis, Kanran Zhou, Paul Felgentreff, Nils Meyer, Andreas Weber, and Harald Sack**. 2019. *Linked Stage Graph.* In SEMANTICS Posters&Demos.