Linked Stage Graph

Tabea Tietz[1,2], Jörg Waitelonis[3], Kanran Zhou[1a], Paul Felgentreff[4], Nils Meyer[5], Andreas Weber[5], and Harald Sack[1,2,3]

[1] Karlsruhe Institute of Technology, Institute AIFB, Germany
[a]kanran.zhou@student.kit.edu
[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
firstname.lastname@fiz-karlsruhe.de
[3] yovisto GmbH, Potsdam, Germany
joerg@yovisto.com
[4] paul.felgentreff@gmail.com
[5] Baden-Württemberg State Archives, Germany
firstname.lastname@la-bw.de

Abstract. Archives today are publishing their cultural heritage data on the Web for exploration. However, for archive novices the traditional archival structures often are not intuitive and difficult to understand, and thus challenges data access and consumption. To tackle this problem, *Linked Stage Graph* was developed, a knowledge graph (KG) on the foundation of historical data about the Stuttgart State Theater. The data was made available by the Baden-Württemberg State Archives for the Coding da Vinci hackathon. This demo paper contributes the KG, a SPARQL endpoint, named entity extraction and linking to existing authoritative KGs as well as a dedicated user interface for exploration.

Keywords: cultural heritage · linked data · knowledge graph · UI.

1 Introduction

Digitizing cultural heritage has been a major task for galleries, libraries, archives and museums (GLAM) for many years now. As a result, a number of Web based platforms have been developed with the goal to enable researchers to access and analyze the data scientifically as well as to allow the general public to explore the data. However, many archival web platforms present their content organized in a way familiar only to archive experts but users who are unfamiliar with archival practise to structure information often find it challenging to access and explore the provided content [1,2].

The Baden-Württemberg State Archives in Germany recognized this issue and opened up their data to the Coding da Vinci initiative, the first German open cultural heritage hackathon. The initiative organizes several hack events a year, bringing together GLAM institutions as data providers and computer scientists, designers, digital humanists to develop creative and interesting applications on the foundation of these cultural heritage data. The Baden-Württemberg State Archives published a dataset about the Stuttgart State Theatres containing

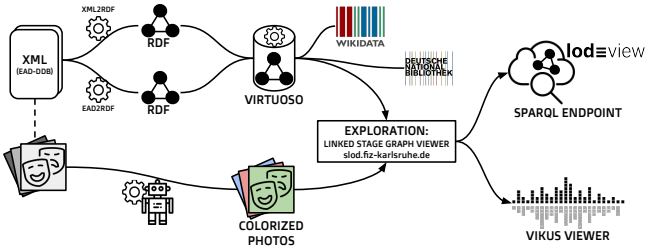

Fig. 1: Workflow and architecture.

7.000 historical black and white photographs along with EAD-XML metadata to be 'hacked' by the creatives. The photographs and metadata cover the period from the 1890s to the 1940s. Especially in Germany, this time period was marked by social and political upheavals in which democracy, freedom of speech and creativity in particular were challenged. For this reason, the data provided are still of enormous relevance today. For instance, the data reveal which theater performances were allowed to be played in these difficult times and how certain characters were displayed [4]. In order to consume and optimally share these data, specific requirements must be met. These include in particular interoperability, availability and comprehensibility on different levels. Linked Data based knowledge graphs (KG) have established as practical means to formally encode, integrate and share data.

In this demo paper *Linked Stage Graph*[6] is presented, a KG developed during the Coding da Vinci Süd 2019 hackathon[7] on the foundation of the aforementioned archival data about the Stuttgart State Theaters. The goal of *Linked Stage Graph* is to enable researchers as well as the general public to access, analyze and explore the data in intuitive, interesting and useful ways. Along with the KG, the presented prototype demo contributes a publicly available SPARQL endpoint[8] to enable sophisticated queries for expert users, the extraction and linking of named entities mentioned in the metadata to the Wikidata KG and the German Integrated Authority File (GND)[9], a timeline interface for data exploration and lessons learned. As an additional feature, all 7.000 black and white photographs were colorized using open source tools based on machine learning (ML). All code of this demo is published and freely available on GitHub[10].

2 Linked Stage Graph

This section presents the main contributions of the demo, *Linked Stage Graph*.

[6] http://slod.fiz-karlsruhe.de/
[7] https://codingdavinci.de/events/sued/
[8] http://slod.fiz-karlsruhe.de/sparql
[9] https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html
[10] https://github.com/ISE-FIZKarlsruhe/LinkedStageGraph

2.1 Dataset

The project is based on an archival fonds of around 7.000 historical photographs from the Stuttgart State Theaters. The photographs depict scenes and characters from a wide range of productions from opera to childrens theater dating from the 1890s to the 1940s. In 2009, the Ludwigsburg State Archives took the fonds into their custody. It consisted of 572 sleeves containing prints, photographic plates, nitrate films, photographic negatives and positive images. The archival description captures (where possible) the title and author of the play, the directors, choreographers and designers of each production. The provided dataset consists of JPEGs and an EAD-XML file (cf. Fig.1 left). The Encoded Archival Description[11] (EAD) is a documentary XML standard for the description of archival finding aids maintained by the Technical Subcommittee for Encoded Archival Description of the Society of American Archivists together with the Library of Congress. For users unfamiliar with an archives content structure it is difficult interacting with EAD encoded finding aids, because often the archive's hierarchy has to be navigated to extract meaningful information [1,2]. Also, it has been widely recognized that EAD complicates user interaction with the data because operations like accessing a specific item on-the-fly are either impossible or inefficient. Automated processing is another issue since the degree of freedom for expressing information within EAD is too high [6]. This attempts to overcome these shortcomings by transforming the EAD-XML into an RDF representation.

2.2 Knowledge Graph

Fig. 1 presents the workflow. Before further processing the input XML file, some adaptions were necessary like adapting the XML name spaces, adding repository code and language code attributes to the EAD unitid XML-tags and replacing the lb tag with XML entity
 . The actual XML to RDF conversion was performed with two different approaches. First, the generic ReDeFer XML2RDF[12] converter was used. The second approach applied the EAD2RDF[13] XSLT stylesheet to transform the provided XML to RDF. As expected, both methods produced different results. While the EAD2RDF stylesheet result were incomplete (e.g. did not manage to transform the information about the image files), the XML2RDF converter produced a vast amount of blank nodes, which are difficult to query and to navigate, but, the results were more complete. Both methods created different IRIs to identify the actual subject of interest: the archival unit. While XML2RDF preferred the archival identifier[14] the EAD2RDF transformation created the IRIs from the EAD unitid[15]. From a computer science

[11] `https://www.loc.gov/ead/`

[12] `http://rhizomik.net/html/redefer/`

[13] `http://data.archiveshub.ac.uk/ead2rdf/`

[14] E.g. `http://slod.fiz-karlsruhe.de/labw-2-2599382`

[15] E.g. from 'Abt. Staatsarchiv Ludwigsburg, E 18 III Nr 6' `http://slod.fiz-karlsruhe.de/id/archivalresource/abt.staatsarchivludwigsburg,e18iiibu161`

perspective the first type of IRI was considered more stable, and that's why it was chosen as identifier. The results were merged through mapping the archival unit titles and the archival identifiers. Finally, the unwanted IRIs were removed. Many literals in the dataset contained unstructured information, i.e. the titles included also a play's author name and the abstracts contained further information about involved persons and roles. To extract this information a script was created. This also involved to define the vocabulary to model the persons types of contribution and roles. The aim was to reuse existing vocabularies as best as possible. Due to the clear definition of the domain the ambiguity in the data was rather low. This enabled to map plays, person, and location names to Wikidata and GND very quickly. Therefore, a dictionary of potentially relevant resources form the vocabularies was extracted and an exact string matching was performed. Finally, all information was deployed to an instance of the Virtuoso triple store also providing a SPARQL endpoint.

2.3 Exploration

A variety of visual means was implemented and utilized to explore the archive data. To bring the historical black and white images to life, they were automatically colorized using an open source tool based on ML [7]. To oversee the RDF data in a table view, an instance of the open source LodView RDF viewer was adapted and deployed [5]. The *Linked Stage Graph Viewer*[16] was implemented for this use case. The viewer is shown in Fig 2. It presents a timeline visualization with the goal to let the user explore the rich and detailed images in a more prominent way without too complex means of interaction to reduce the technical barriers of engaging with the content. The user can scroll through the images with an overview of the timeline on the right ①. One large representative image for each performance is shown ② with further thumbnails on the bottom ③. Swiping left or right reveals further plays which took place during the same year ④. Clicking on images will direct the user to the Lodview interface and reveal all data available for the play. Next to the implemented *Linked Stage Graph Viewer*, the *Vikus Viewer*[17] was utilized and connected to the dataset as well. The viewer was previously developed by [3] and enables intuitive content exploration in a timeline view as well as search and content clustering. During the demo session, all described interfaces can be used to explore the dataset.

3 Conclusion

In this paper *Linked Stage Graph* is presented, a KG based on historical data about the Stuttgart State Theater. Next to the KG, a SPARQL endpoint was released, named entities mentioned in the metadata were extracted and linked to existing KGs and a user interface was developed. The demo was created during

[16] `http://slod.fiz-karlsruhe.de/#Viewer`
[17] `http://slod.fiz-karlsruhe.de/vikus/`

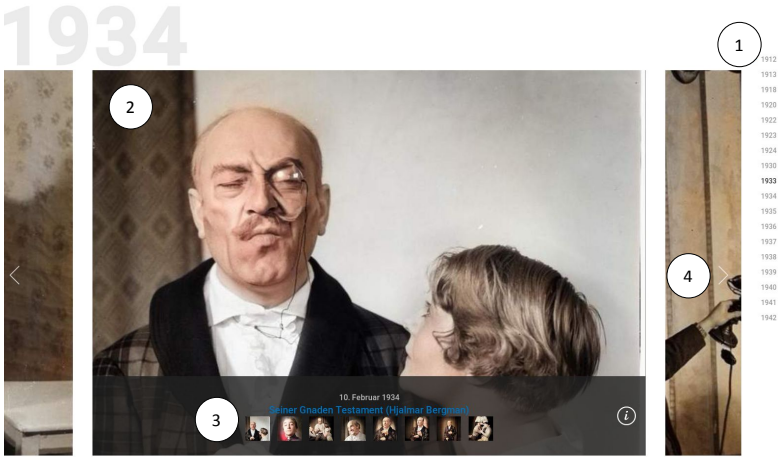

Fig. 2: Linked Stage Graph Viewer

the Coding da Vinci Süd hackathon and was awarded the prize for the "most useful" application. Even though the presented demo implements a use case on theater data only, it is generalizable to a broad range of archive domains since the EAD-XML format is widely used in archives. Future work will involve the improvement of the underlying ontology, a more exhaustive entity linking and further means of visual exploration beyond timelines.

Acknowledgement. We would like to thank Coding da Vinci for connecting cultural institutions with creatives to develop innovative applications.

References

1. Ferro, N., Silvello, G.: From users to systems: Identifying and overcoming barriers to efficiently access archival data. In: ACHS@ JCDL (2016)
2. Freund, L., Toms, E.G.: Interacting with archival finding aids. Journal of the Association for Information Science and Technology **67**(4), 994–1008 (2016)
3. Glinka, K., Dörk, M.: Museum im display. visualisierung kultureller sammlungen (vikus). Konferenzband zur 22. Berliner Veranstaltung der internationalen EVA-Serie: Electronic Media and Visual Arts 2015 (2015)
4. Halbach, F.: Judenrollen: Darstellungsformen im europäischen Theater von der Restauration bis zur Zwischenkriegszeit, vol. 70. Walter de Gruyter (2008)
5. Hildebrandt, F., Pohlmann, A., Omran, H.: Lodview: a computer program for the graphical evaluation of lod score results in exclusion mapping of human disease genes. Computers and biomedical research **26**(6), 592–599 (1993)
6. Prom, C., Rishel, C., Schwartz, S., Fox, K.: A unified platform for archival description and access. In: Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007. pp. 157–166 (2007)
7. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics (TOG) **9**(4) (2017)