# Best practices within Digital Humanities to visualize large collections of cultural heritage: Exploring Music Collections

*Emilė Mikutaitė (2784212) , María Estrella Puertollano (2781167) and Yihyun Kim (2787967)*
Vrije Universiteit Amsterdam

**Abstract.** Cultural heritage organisations possess more digitalized data than ever, and making their data accessible comes with challenges in their structure and usability for researchers. With the aim of making linked data accessible and readable, the Netherlands Institute for Sound and Vision (NISV) has created an extensive concert collection MOZ (Muziekopnamen Zendgemachtigden) which is structured using open linked data. We have examined this collection through SPARQL queries, Python notebooks and available documentation to create a schema for the data structure that will serve as a base to create a generous interface used to explore music collections. With this schema in mind we reviewed the best practices for visualizing linked data in digital humanities and how they can be applied to the MOZ collection. These contributions will make the MOZ dataset more understandable to its future users, as well as contribute to the creation of its data story.

**Keywords:** Music collection, Linked data, Digital Humanities, Visualization

## 1. Introduction

The research project, called “Exploring Music Collections through data stories, exploratory interfaces and innovative applications” has been started with the aim of obtaining a readable visualization for the large music collection that the NISV counts with. This music collection contains concerts performed by a large selection of people, from Beethoven to Metallica, englobing a considerable period of time, and has a rich amount of Dutch artists. Therefore, it can be noted that the variability of the dataset is ample, thus, making the categorization difficult. However, when it comes to data storytelling, the key is to faithfully plot the data features, together creating an exploratory idea of the dataset. From this perspective, we will explore what makes up a data story in Digital Humanities, and more specifically music collections.

Nevertheless, this work could only be done after having created a theoretical framework on linked data, taking into consideration that the nature of the data is N-triples (subject linked to predicate, and this linked to object). For this, several digital tools will be needed to understand the data structure and make use of the data. This consequently places the research in the computer science field as well because not only is programming needed to carry out this research, but also computational thinking to fully understand how the dataset has been created and which approach must be taken to make the most of the available information.

Considering that, the aim of the study is to offer the best practices within the Digital Humanities field to visualize (computer science) large collections of cultural heritage data (humanities and social sciences). This specific goal of our research is part of a bigger goal for the NISV. Our approach and contribution will definitely be useful in that it will serve as a basis to create a generous interface for exploring music collections, aimed by the Institute. As in this project we have established the data structure, and obtained a literature review, the NISV will be able to make good use of these to put the proposed visualizations and tools into practice for their storytelling of the data. However, not only will our project be of use for the specifics of the NISV, but it will also contribute scientifically to develop new tools based on the necessary data and the theoretical specifications we studied for Digital Humanities works.

## 2. Problem statement

Visualizing large data is a well-known challenge in the field of digital humanities that has received a lot of attention. Additionally, linked data provides more complications, as the relationships between data objects and properties are very extensive and often surpass the limit of a singular dataset. The computational challenge in visualizing such data is that it is stored in triples which are more difficult to read for more traditional data visualization tools, and requires either extracting the data via additional code to structure it in a tabular way or using specialized tools for linked data. Therefore, we formulated the following research question and subquestions:

*RQ:* What are the best ways to visualize large linked cultural heritage data?

*SubRQs:*

- What are the common practices in visualizing linked data within digital humanities?
- How can a schema for linked data be visualized?
- What makes up a data story for cultural heritage linked data?

## 3. Dataset

For this project, we worked with the MOZ data, namely Muziekopnamen Zendgemachtigden, which means *music recordings broadcasters*. The dataset contains the original concert and studio recordings created by public broadcasters. To facilitate users of this dataset, NISV mapped relevant metadata fields from their internal metadata format to Schema.org. Within the dataset, the Schema.org vocabulary is used to cover entities, relationships between entities and actions. For example, the concerts, which themselves are the type *sdo:CreativeWork* ('sdo' stands for Schema Dot Org) in the MOZ data are linked to a *sdo:CreativeWorkSeason* using the property *sdo:partOfSeason*. Moreover, the dataset offers a SPARQL endpoint, which allows the users to build their own queries or use and modify sample queries in the user interface.

## 4. Related Work

When analyzing the state of the art, we encounter large amounts of literature concerning visualizing historical data, for example, but very little about music collections. It seems evident that there are huge differences between these two fields; however, both belong to the Humanities field. Therefore, we have been able to profit from every type of literature within Humanities. More accurately, our literature is based on the state of the field of the Digital Humanities, englobing both computational and humanities aspects.

## 4.1 Computational domain

As mentioned in previous sections, one of the intricacies of the project is having linked data as a dataset. We can define linked data as the data that "involves creating identifiers for things or resources on the Web and then linking these resources together, using statements in a standard format called RDF (Resource Description Framework)" (Neish, 2014).

We have all made use of hyperlinks when browsing the Internet, because most of the data is linked these days. For it to be linked, some relationships are established between the objects, and these of our dataset are the ones made explicit in the data schema offered in the sections below. The relationships follow a semantic criterion in the N-triples: subject, linked to predicate, linked to object. The abundance of these N-triples has lead to "schema.org" to exist. This web shows the possible existing objects next to a classification of them, as well as the relationships that could be established between them. Schema.org is a standard for this semantic content.

From the definition offered in the first paragraph, we can deduce that linked data is not a compact set of data, but a wide variety of possibilities to explore through the links. Therefore, and focusing on another of the project's branches, we introduce the concept of "generous interfaces". These consist of exploratory interfaces for the user to have a general overview of the search of interest before accessing a specific website. They are the solution to the common interfaces, that hold back information through the search bar, for instance. Generous interfaces need to be created from the computational field making use of tools such as *Observable* (recommended by our problem owners and used by them in other projects), based on JavaScript programming. In our case, we have not been able to test it because the complete dataset is still not extractable. However, there are some examples, such as the one shown below, in which we observe the generous interface enabling the filtering of the data and the exploratory nature at the same time.

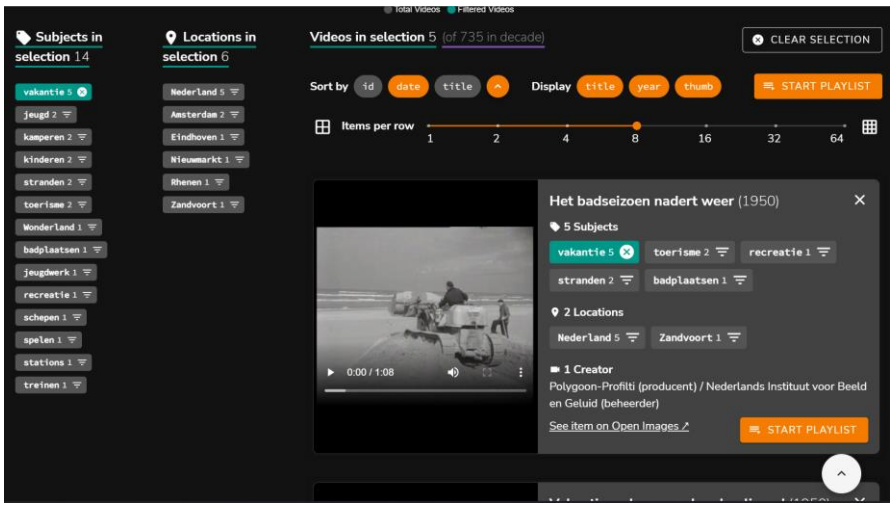



*Figure 1. Open Images Browser generous interface.*

## 4.2 Humanities domain

Data narrative concerns this project to the extent that there is a dataset that needs to be explained to the user in an interactive way. Therefore, we have looked for other projects that specify the best practices for storytelling, and this is when Humanities comes clearly into play.

*CLARIAH Media Suite* (CLARIAH Media Suite - Home, n.d.) offers tools on how to best create a data story; it is a platform that makes Dutch archives available to be used for projects such as ours. It can be deduced

that the most common practice is using a bar graph with a hover function, which in our case does not fit because of the lack of exploratory nature in that type of graphs. However, it has been still interesting to understand the common approach in some of the projects, which is to show statistics in an interactive way, at the same time that the dataset is explained in terms of content: variables, attributes and relationships.

### 4.3 Digital Humanities domain

This section is included to expose the results on our search for the best practices for visualizing large collections within Digital Humanities. There is a clear division between temporal and non-temporal graphs. In our case, it has been interesting to investigate both, taking into account that time is a present variable in our dataset, although not the most representative. A table unifying this content can be found in Appendix A.

We have also observed some challenges in other projects using these graphs; for example, creating an exploratory visualization based on time. Moreover, as much as maps are always a useful tool, they seem to focus only on the historical data that, as said, has been useful to think of the approach, but not as a source of ideas for visualization.

Finally, it can be noted that the grid mosaic idea is a common practice when thinking about generous interfaces, useful because it enables both an informative and exploratory visualization that can serve as a basis to the final interface.

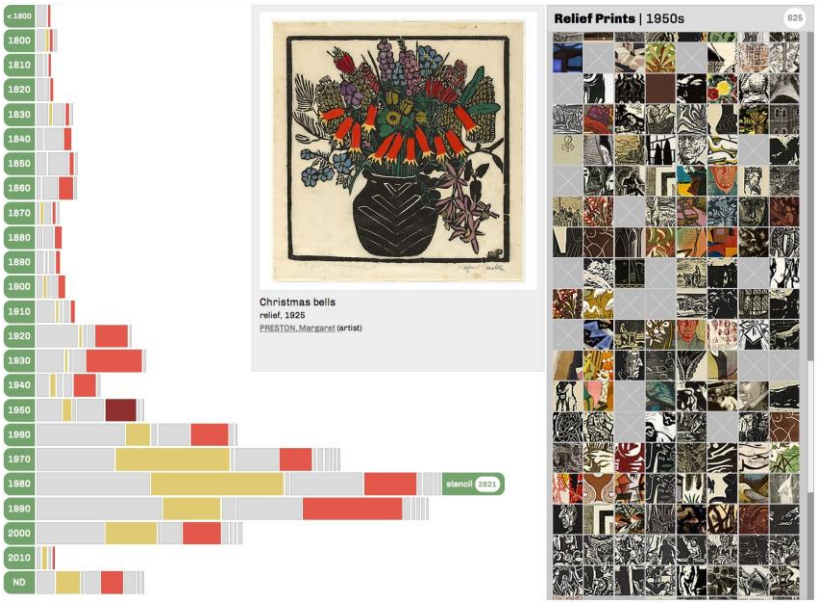



*Figure 2. Generous interface based on exploratory mosaics for the Australian Prints and Printmaking.*

## 5. Method/approach

### 5.1 Research setup and data acquisition

The project is primarily approached from two angles: a literature review on large (linked data) visualization practices in digital humanities, and using the available dataset to improve upon its existing data story. By having descriptions and examples of the best visualization practices we could then move on to drafting possible solutions for representing the MOZ dataset. Extracting the data was a challenge, as it needed to be acquired through queries using a SPARQL endpoint, which we did not have prior experience with. Also, it could not be

downloaded directly from the SPARQL endpoint, and instead only acquired via writing additional code using a Python notebook. We were assisted with this issue by the problem owners, who provided us the example code and a small subset of the data.

In the process of accessing and reviewing the data, we uncovered that the dataset itself needs more attention, as its structure was difficult to uncover. It currently does not have a complete documentation available to the public and lacks an overall description, including its limitations. Therefore, creating a complete schema gained priority, as it needs clear visualization before extracting any specific data for content visualization.

### 5.2 Solution approach and design

After evaluating the initial challenges that came up when navigating the project, several objectives were formulated. Firstly, we need to review the best practices of visualizing linked data and based on the outcome propose a solution for the MOZ dataset. For that, a literature review must be conducted, as well as an overview of available exampled. Secondly, a complete schema for the dataset must be compiled and visualized accordingly. That is done by compiling all accessible sources relating to the data, such as a description that was provided by the problem owners, an incomplete schema, websites of the specific entries in the dataset (a concert, an artist and an institution), as well as employing additional tools such as GraphDB to read the available data subset. Finally, any other necessary aspects of the data story should be identified, and if possible formulated to complement the primary visualization and dataset schema.

### 5.3 Evaluation setup

The evaluation process is parallel to the goals that have been established for this project. Each goal is evaluated separately by two evaluation criteria: completeness and added value. Completeness describes which stage of the specific goal was reached, for example whether we provided existing visualization examples that could be used in the future, described their adaptation to the case, made sketches, 2D versions or fully adapted interactive visualizations. Added value is a reflection on how the specific solution is useful in the context of digital humanities, for the problem owners and also for future users of the dataset. Finally, the goals should also be reviewed in combination by how well they compliment each other and by whether together they address the main issues that were voiced by the problem owners and also uncovered during the process. The feedback that is received from the problem owners can also be considered as additional external evaluation of the project.

## 6. Results

### 6.1 Exploratory dataset visualization

The first objective in creating a data story for the MOZ dataset was providing an interactive exploratory visualization which would allow the user to peruse the dataset in a more intuitive way. A 2D sketch of such visualization can be seen in Figure 3. The idea for this initial layout comes from grid mosaics, which are usually made up of either individual collection entries or their categories. We decided on the latter, as being able to overview all singular concerts could prove more overwhelming, rather than aid comprehension. Meanwhile,

Providing some categorization of the entities can give data users a good idea of what kind of related information they would be able to find in the dataset.


*Figure 3. Exploratory visualization for the MOZ data story.*

Due to difficuties with accessing and analysing data, this visualization serves as a very general suggestion that should be further refined to match the scope and content of the dataset more precisely. In this case, we decided to use concerts as a central subject and highlight a few prominent object classes and properties, such as location, creative work season or date. These categories are not the end goal - they serve as an entryway to a more detailed and interactive interface.

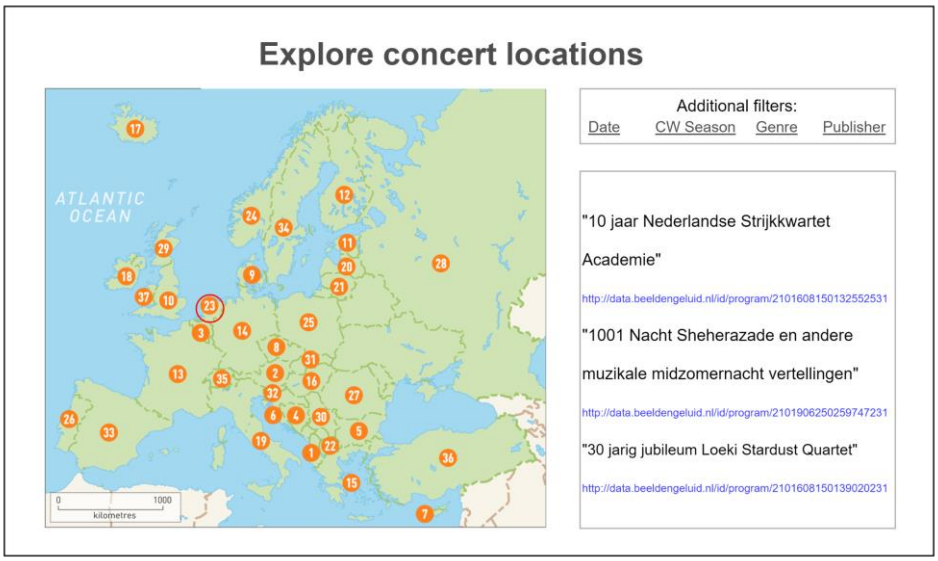



*Figure 4. Generous interface example for concert location data.*

A 2D example of such an interactive interface can be seen in Figure 4. This illustration shows a mock version of an exploratory interface that a user is presented with when they click on ‘Explore concert locations’ in the initial grid. The idea is based on the concept of generous interfaces; it aims to have an exploratory section (i.e., interactive map), which the user can zoom in and look at, as well as a filtering system which can be used to refine their search and look up concerts that match the specific criteria. Similar interactive interfaces with additional

filtering could be created for the other categories, such as timelines for different time periods, as well as visualizing artists active throughout the years or bubble charts for genre and publisher overviews.

## 6.2 Data schema visualization

Navigating the data, especially through SPARQL inquiries, proved problematic due to the lack of a complete dataset description. Therefore, one of the objectives that emerged during this project was to create a complete data schema. Due to the nature of linked data, portraying its structure in writing is often confusing, as the object relationships are less obvious. Therefore, it became important to visualize the linked data structure, as it allows for more clarity and user-friendly exploration. The ideal version of the data schema in this case would be a complete and interactive visualization that can be presented on the MOZ data story website for people who are interested in using the dataset.

The sketches for the data schema are seen in Figures 5 and 6, they were made using Draw.io. The schema consists of the different classes, object properties and data properties, as well as the relationships between them. The data schema has two versions - one is collapsed and only includes the classes (Figure 5), while the other contains all details that could be mapped at this moment (Figure 6). The reasoning for making two versions was to have a clear idea for how the interactive visualization could be designed. Sadly, we were not able to design the final interactive version of the data scheme due to time restrictions and available tools. Still, these sketches contain the most crucial information about the data structure that can inform any future attempts at making a complete schema.

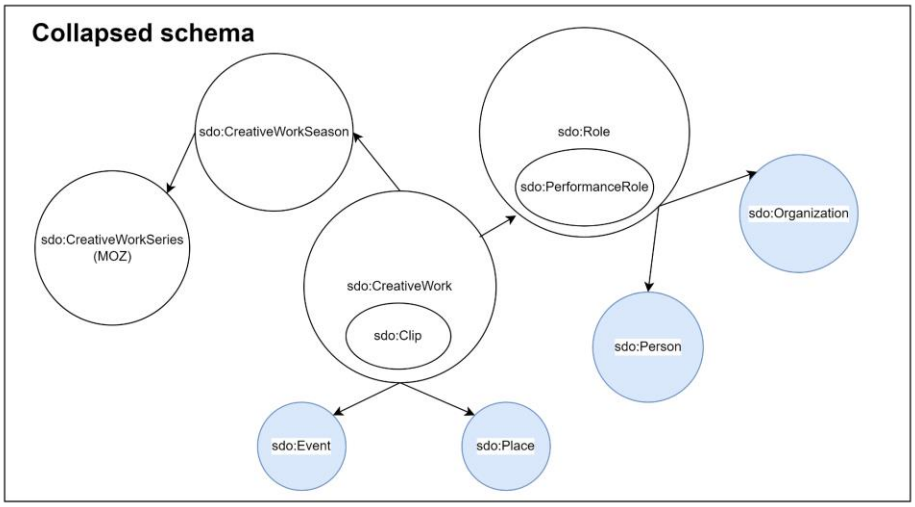



*Figure 5. Collapsed data schema visualization.*

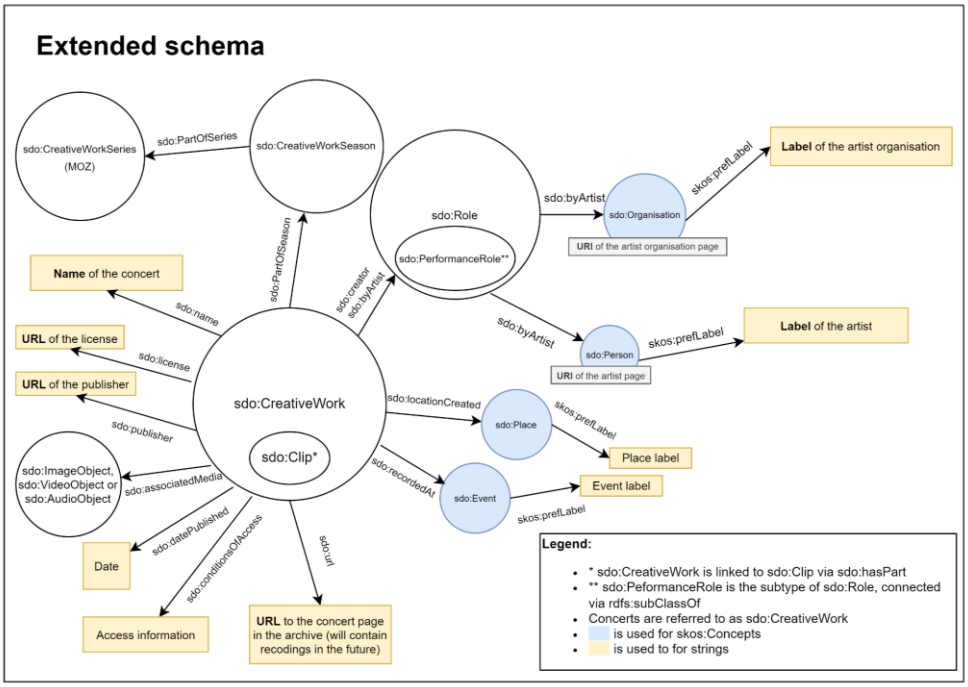



*Figure 6. Extended data schema visualization.*

Although it was not created for the study, we also considered ways to make the schema interactive in the future. Firstly, only the collapsed schema (Figure 5) would be visible to the user. They could reveal connections to other objects by clicking on the class of interest until all the extensions are revealed (Figure 6). Additionally, hovering over an object would provide a small text box with a description and example of the object. Similarly, hovering over an arrow would reveal its property name. There could also be an toggle to show all or none of the property names, depending on the user's wishes.

### 6.3 Dataset disclaimers for future users

As our goal is to make the MOZ dataset more understandable and easier to navigate for new users, we would also like to present several disclaimers that should be displayed under the data schema. By doing so, the future users of this dataset can be proactively informed about the data before accessing and using it. Thus, it minimizes the difficulties they may face when using the data, and also the amount of questions that will be directed to the data owners.

We have therefore compiled a short list of disclaimers that can be found in Appendix B. They are based on the limitations that we have uncovered by trying to access the dataset and make use of it.

## 7. Discussion and Conclusion

### 7.1 Conclusions

In conclusion, this project in collaboration with NISV focused on how to visualize large linked cultural heritage data, which includes creating exploratory visualizations based on generous interfaces and cultural heritage data visualization practices, as well as creating linked data schema visualisations. The resulting visualizations were based on a literature review which gave insight into the best visualization practices within Digital Humanities regarding large-scale data visualization and generous interfaces. While the visualizations are

only example sketches, they provide ideas for how an exploratory interface could be created for the MOZ data story in a way that is helpful and easy to navigate for future users. Moreover, we have created an extensive data schema visualization, from all currently available sources. Although the data schema visualization is not fully complete or interactive either, we have succeeded in creating a good overview of the MOZ dataset, especially its almost prominent elements. Finally, being the first external users the MOZ dataset allowed us to provide NISV with insights about how future users might perceive the dataset, and highlighted what additional information should be included in the data story. To sum up, the resulting visualizations did not only help us get a clear idea of the MOZ dataset but also will allow the problem owner to create the MOZ data story that meet the needs of future users.

## 7.2 Discussion

Having analyzed our own performance in this research project, we consider it important to mention the limitations that we have encountered to enable a better understanding of this process for future projects. Firstly, the dataset consists of linked data structured in N-triples, which we were not previously experienced in. This has not only entailed hard work of learning its structure and intricacies, but also prevented us from working with (or at least successfully using) a lot of digital tools. Additionally, the SPARQL endpoint does not allow to directly download the data, thus the provided subset that was used alongside several other sources made data comprehension more problematic, and might not have been fully representative. However, we did receive some guidance from the problem owners that led us to creating the preliminary data structure schema, which as we have already mentioned is one of the biggest outcomes of this project. At the same time, the dataset is continuously updated, meaning that the provided schema cannot be fully complete.

All of this can be used to make suggestions and open paths for future work. Firstly, additional research should be done in the topic of music (collection) visualizations, as a lot of cultural heritage data revolves around physical objects or images, while more abstract objects have received less attention. Also, visualization of linked data is another topic to further explore in research, as linked data is used increasingly more often but its structure includes many complexities that differ from data structures that more humanities and social science researchers are more familiar with.

When it comes to more practical work, the current project results will need to be further built upon by NISV to create the complete MOZ data story. In this sense, our contribution serves as an initial reference for creating an exploratory interface and visualizing the data structure that will make up a complete data story that meets the need of its users. Additionally, us having been the first external users of the dataset already proved to be useful, as said by the problem owners, because we pointed out some assumptions that need to be made explicit for future data users. After the relevant changes to the MOZ data story are introduced, we would suggest to present it to a few other specialists in the field who have not used the data before. This would allow for additional feedback regarding the user experience and whether the data story serves the needs for cultural heritage professionals and researchers who want to use the data.

When considering our fields of concern, we believe this paper is a considerable contribution to the topic of visualization in cultural heritage studies that, as we mentioned before, tend to focus on interactive maps due to the abundance of historical data. Therefore, documenting out experience working with music-related data will enable other interested researchers to have a better theoretical understanding about the issue, as well as some

visualization ideas both for an exploratory interface and linked dataset. We would also like to note that, from a computational perspective, this paper has shown that digital tools are less accessible for linked data and manual work is still necessary at times; therefore, we suggest exploring existing tools and their capabilities with the aim of making linked data approaches more user-friendly.

## 8. References


*All stories*. (n.d.-a). Media Suite Data Stories. https://mediasuitedatastories.clariah.nl/

*All stories*. (n.d.-b). Media Suite Data Stories. https://mediasuitedatastories.clariah.nl/

*Beeld & Geluid - Sound & Vision Â· Labs*. (n.d.-a). https://labs.beeldengeluid.nl/datasets/muziekopnamen-zendgemachtigden

*Beeld & Geluid - Sound & Vision Â· Labs*. (n.d.-b). https://labsbeeldengeluidnl-git-moz-data-story-beeldengeluid.vercel.app/nl/blogs/moz-dataset-blog

*CLARIAH Media Suite - Home*. (n.d.). https://mediasuite.clariah.nl/

*Comunica Web client*. (n.d.). https://cat.apis.beeldengeluid.nl/

DataJournalism.com. (2021, June 28). *Telling data stories with music*. https://datajournalism.com/read/longreads/data-sonification

*Help:About data - Wikidata*. (n.d.). https://www.wikidata.org/wiki/Help:About_data

Knees, P., Schedl, M., Pohle, T., & Widmer, G. (2006). An innovative three-dimensional user interface for exploring music collections enriched. *Proceedings of the 14th ACM International Conference on Multimedia*. https://doi.org/10.1145/1180639.1180652

*Linked Data for Digital Humanities | Victor de Boer*. (n.d.). http://www.victordeboer.com/linked-data-for-digital-humanities/

Maps, E. S. (n.d.). *Story Maps and the Digital Humanities*. Esri. https://collections.storymaps.esri.com/humanities/

Pampalk, E., Dixon, S., & Widmer, G. (2004). Exploring Music Collections by Browsing Different Views. *Computer Music Journal*, *28*(2), 49–62. https://doi.org/10.1162/014892604323112248

*RDF 1.1 N-Triples*. (2014, February 25). https://www.w3.org/TR/n-triples/

*The Sensory Moving Image Archive*. (n.d.). https://bertspaan.nl/semia/

*Visualising Cultural Heritage Data | Blog*. (n.d.). Data Visualisation Hub - the University of Sheffield. https://dataviz.shef.ac.uk/blog/15/06/2021/Visualising-Cultural-Heritage-Data/

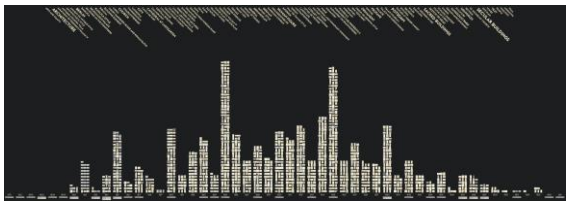
Windhager, F., Federico, P., Schreder, G., Glinka, K., Dork, M., Miksch, S., & Mayr, E. (2019). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics*, *25*(6), 2311–2330. https://doi.org/10.1109/tvcg.2018.2830759

Neish, P. (2014). Linked data: what is it and why should you care?[a]. *The Australian Library Journal*, *64*(1), 3–10. https://doi.org/10.1080/00049670.2014.974004

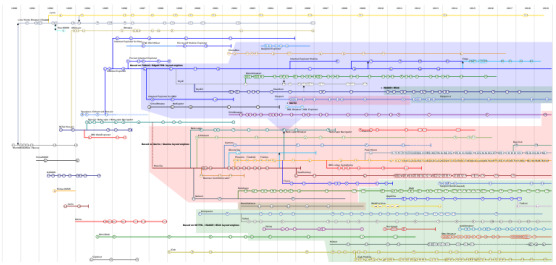
# Appendix A. Visualizing large cultural heritage data.

## Temporal graphs

1. Time axis in 2D



2. Timeline in 1D

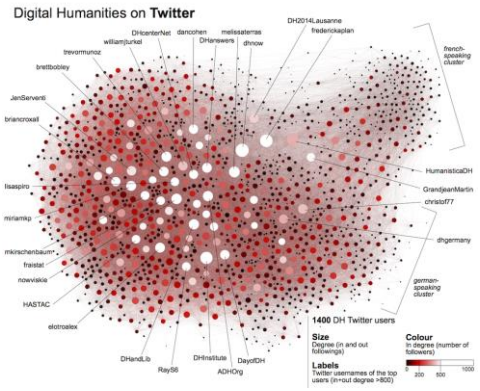


## Non-temporal graphs

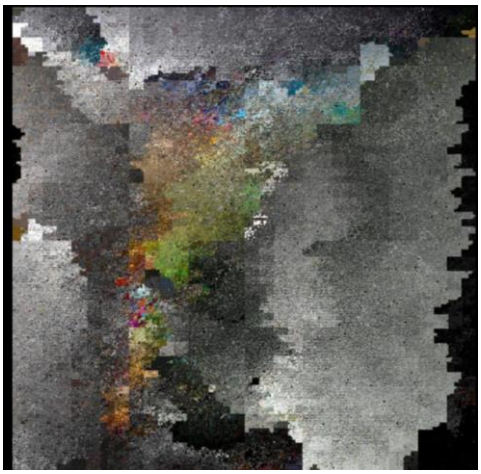
1. List of slideshows

2. Maps

3. Networks




4. Grid mosaic



## Appendix B. Dataset disclaimers for future users

Before trying to access the data, please be aware of the following:

- At the moment, the SPARQL endpoint does not allow the users to directly download the dataset or its subsets.
- Please consult the data schema before writing SPARQL queries for exact object and predicate names.
- More specific SPARQL query examples for this dataset can be found at the endpoint, under “Type or pick a query” (i.e., MOZ Concerts - top 10 artists).