Group Assignment 1: Project Characterization

Interactive-Visual Data Analysis, Fall 2024

Dionigi Rodriguez (24-755-688)   **Cyril Smetanka (24-754-434)**
Patrick Sproll (19-733-104)

November 9, 2025

Domain

We have set out to make an interactive surface for Greta Franzini's manually collected "Catalogue of Digital Editions", which consists of 350 works from digital humanities projects. Our aim is to enable analysis of these works in relation to each other and to provide an interface that will primarily be used for search actions, mainly exploration, while also providing features for consumption actions like discovery and enjoy.

The catalogue has a website already (https://dig-ed-cat.acdh.oeaw.ac.at/editions.html), but it is very barebones when it comes to visual exploration, providing only a map of the funding institutions, which doesn't enable exploration in the way that domain experts would ideally navigate this data, as it is missing any grouping of eras, languages, or types of audiences the works are catered towards. In the course of our study we will follow and refer to Tamara Munzner nomeclature [1].

What

The catalogue consists of a list of 350 works from digital humanities projects with a total of 52 attributes each (including an ID per item), obtained in tabular form as a CSV file from the following GitHub repository (Link)

Of the 52 attributes, 25 are encoded as dummy variables, 15 as text, 8 as categorical variables, and 4 as numerical variables., of which 3 are dates/years. Categorical variables allow the editions to be more finely described, again by inherent attribute e.g. the historical period. Works range from the antiquity to contemporary, with "Middle Ages", " Long Nineteenth Century" and " Early Modern" dominating. Predominantly, editions are in either English, Latin or German, which comprise more than half of the catalogue.

Of the 18'252 cells, only 8 have missing values, in just three columns. Given the data was prepared by an experienced researcher, we consider it both reliable and clean, requiring little preprocessing except the handling of the missing values, which will be done through replacement. We intend to augment the dataset with a number of LLM generated attributes to facilitate exploration and assessment of works by researchers through similarities and connections between items: Author school of thought, 5 keywords, authoritativeness, and a statement for the renown of a work as well as a quick work description to provide some detail on the contents of a work during exploration. We intend to use mainly intrinsic attributes for the project, not the technical ones, as the target audience is primarily interested in these aspects.
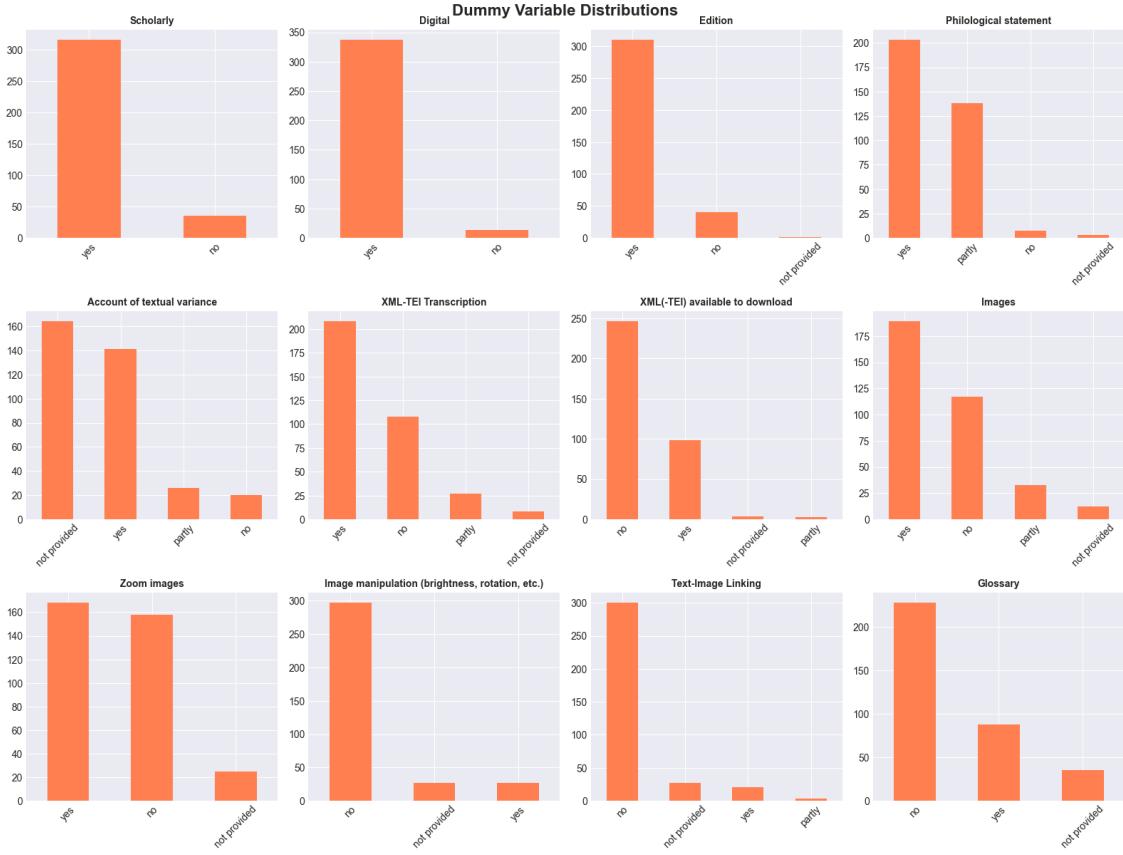
Figure 1: Distribution of dummy variables in the catalogue.

Why

The goal according to the problem statement is to let users (primarily historians, researchers and digital library users) explore the contents of the catalogue, while also allowing them to do a directed search. The task falls squarely into the data exploration and interactive relation discovery research streams. We are considering a sub tool that enables users to tag their own connections between works, which would touch upon interactive data labeling. The following tasks need to be covered:

- Gain overview – Users will need to get a grasp of what they are able to find in the catalogue and if there may be something that fits their needs and interests, focused on intrinsic attributes, not technical ones.

- Identify relationships – We facilitate this by guiding exploration among relationships between works, such as time period, or author. We will use clustering and LLM generated keywords to find similar works.

- Filter by attributes – To reduce noise and narrow down the search space, users need to be able to filter by all attributes, technical and non technical.

- Judge reliability – Users need to be able to assess the quality of the data. We will provide LLM generated attributes such as renown and authoritativeness along side the values contained in the catalogue. We will provide an overall score as well.

- Tag relationships – To record their own train of thought and help organize their research, users need to be able to tag relationships between works.

How

- Gain overview: Timeline with colored background blocks and works plotted, a block of title, author and description of a work, a bar chart of languages, and a cluster plot based on keywords as initial overview, all linked and can be narrowed down to suit any subset of works.

- Identify relationships: A network graph and cluster plot with redundant encoding by colour and shape.

- Filter by attributes: A search function by title or author for specific queries, as well as traditional filter by attribute. These will come in the form of sliders, or drag and drop pills. Where possible, we will also provide histograms.

- Judge reliability: We will provide a score represented as a colored bar, as well as the attributes contributing to the score.

- Tag relationships: A network graph where users can add edges between works to represent their own relationships and tag them as well as organize in groups.
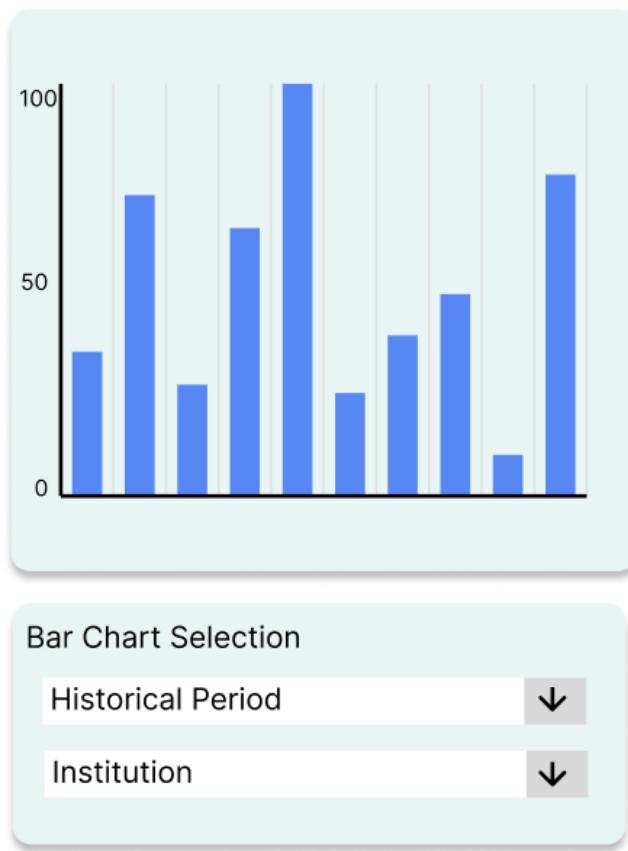

Figure 2: The user defineable bar chart.
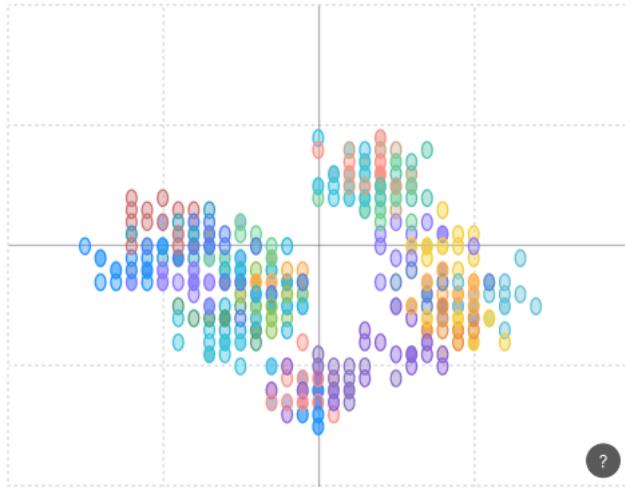
Figure 3: The cluster plot.

Figure 4: The detailed example preview of a single item
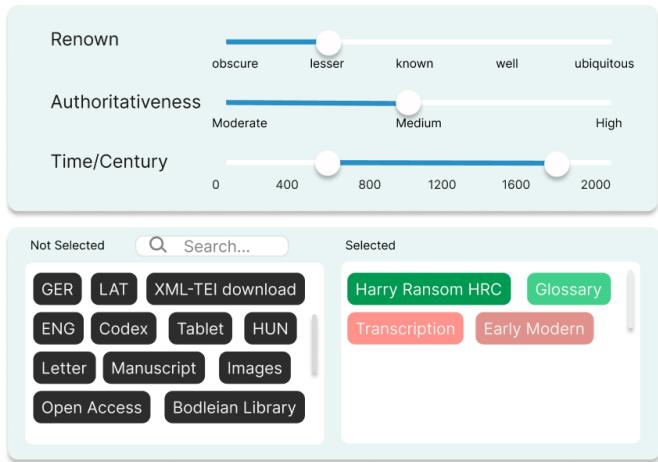
Figure 5: global filter settings component.
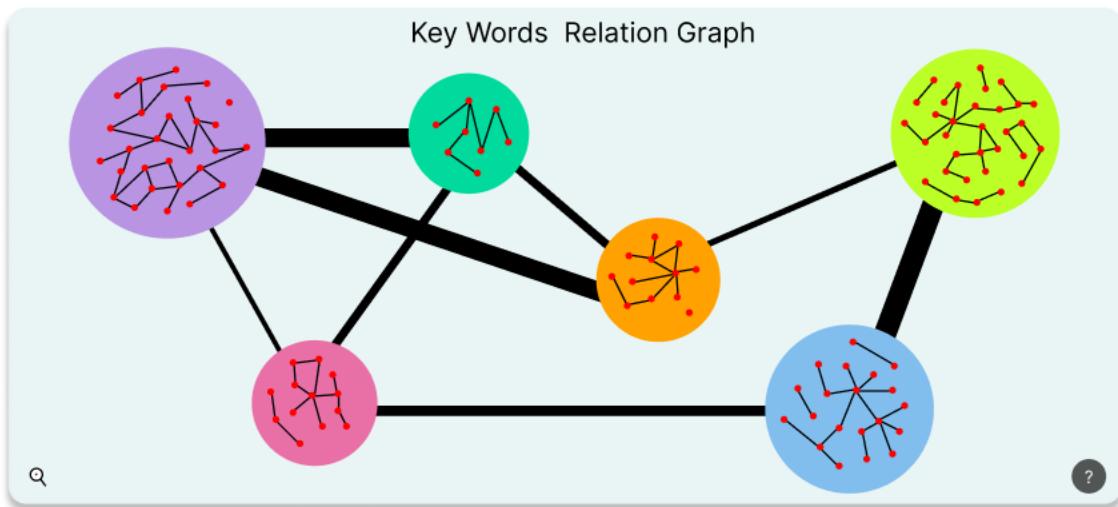
Figure 6: Relation graph
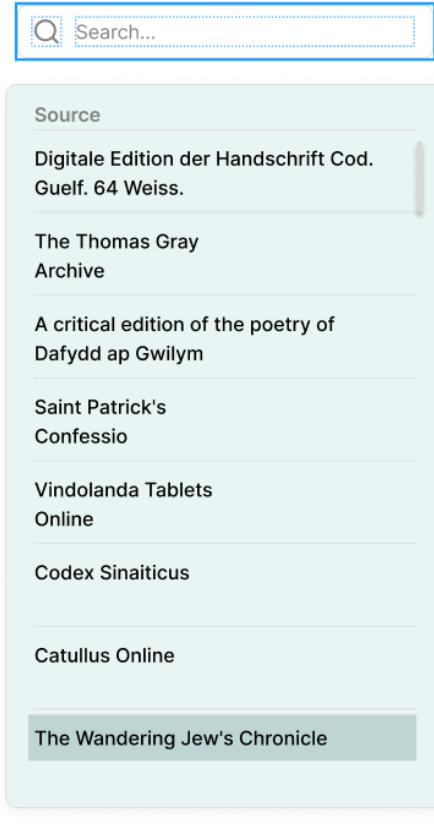
Figure 7: List of works
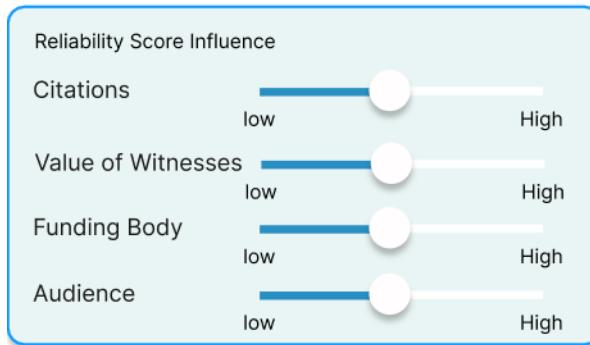
Figure 8: Reliability score settings
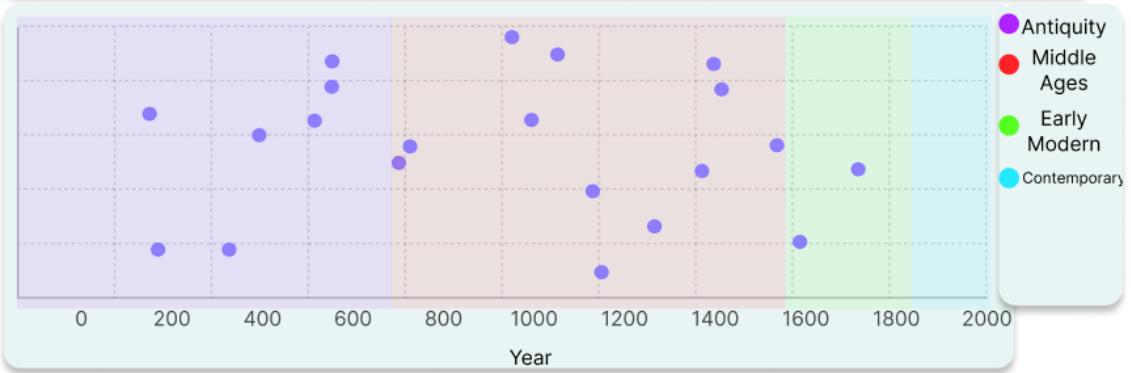
Figure 9: Timeline of works

We will provide this in a single dashboard with linked visualisations as outlined above. We will allow users to define global filters and apply them to individual elements.

We will use two main techniques, the k-means [2] algorithm and PCA [3] (Principal Component Analysis), to identify significant clusters and represent the data in a lower-dimensional space. In addition, we will leverage a pretrained transformer model (e.g., BERT [4]) to generate semantically meaningful embeddings[5] of various attributes, providing a numerical representation of each article for use with the clustering and dimensionality reduction algorithms.
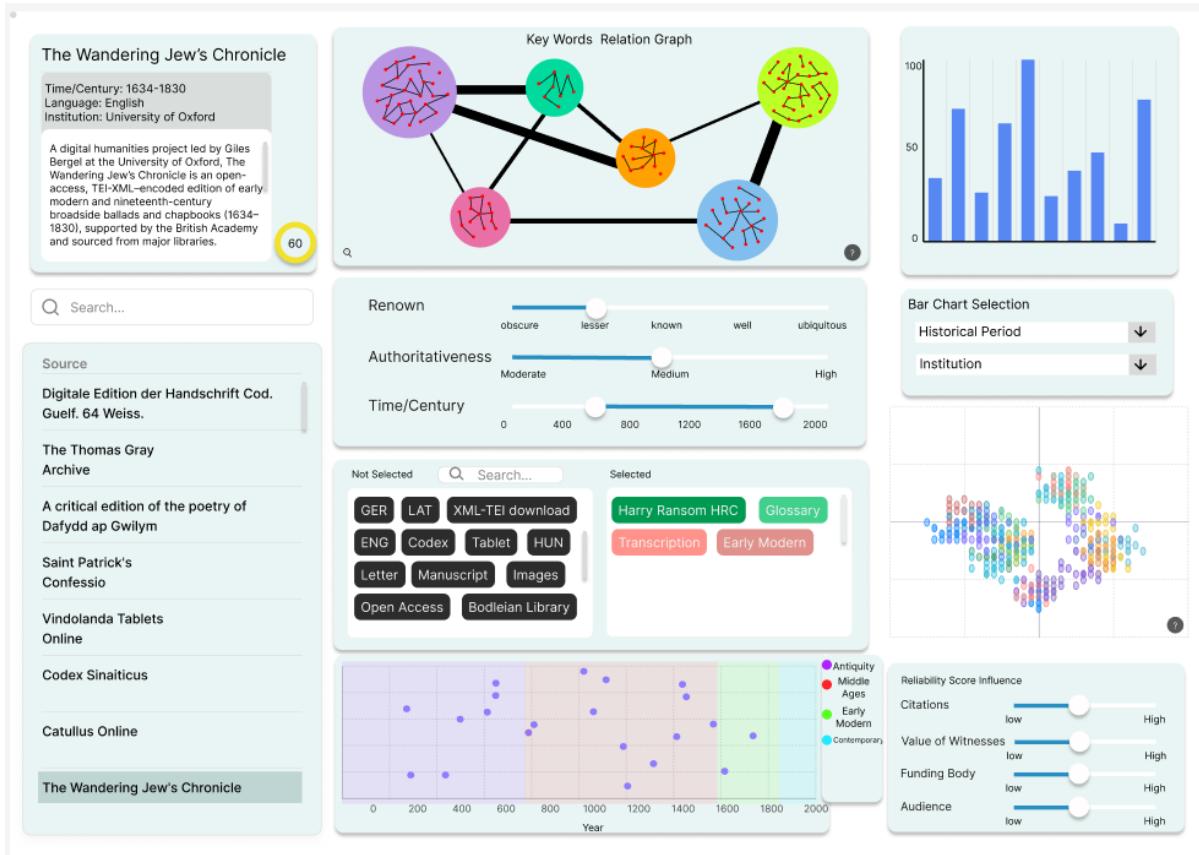
Figure 10: The collected dashboard across all visualisations.

Group Dynamics

Given that our data is fairly clean and well structured already, we have no need for an explicit Data Steward/ Data Shaper role as outlined in the course material. Dionigi will largely take on the roles of data engineer and ML/AI engineer, supported by Cyril, as both study AI. Cyril will also take on the role of a generalist, supporting frontend development and documentation. While we are lacking a research scientist, Patrick will take on the role of technical analyst, and support Cyril in frontend development and documentation as well. All together will serve as evangelists in so far as it is necessary.

References

[1] T. Munzner, *Visualization Analysis and Design*, ser. AK Peters Visualization Series. CRC Press, 2014. [Online]. Available: https://doi.org/10.1201/b17511

[2] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979. [Online]. Available: http://www.jstor.org/stable/2346830

[3] J. Shlens, "A tutorial on principal component analysis," *CoRR*, vol. abs/1404.1100, 2014. [Online]. Available: http://arxiv.org/abs/1404.1100

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] P. Umair Ali Khan, "The power of embeddings for semantic search," *medium.com*, 2024. [Online]. Available: https://medium.com/@umairali.khan/the-power-of-embeddings-for-semantic-search-8883f3fe8ba2

PROJECT DETAILS	PROJECT NAME	DOMAIN	KEY APPLICATION
	IVDA Group Project G1	University Project	Interactive Visual Data Analysis

CONTACT(S)	NAME(S)	EMAIL(S)	AFFILIATION(S)
	Patrick Sproll	patrick.sproll@econ.uzh.ch	University of Zürich (UZH)
	Dionigi Rodriguez	dionigi.rodriguez@uzh.ch	University of Zürich (UZH)
	Cyril Smetanka	cyril.smetanka@uzh.ch	University of Zürich (UZH)

MODEL(S)	MODEL NAME(S)	VERSION(S)	
	Github Copilot	latest	

IDEATION	GENERATING IDEAS, OUTLINES, AND WORKFLOWS	IMPROVING EXISTING IDEAS	FINDING GAPS OR COMPARE ASPECTS OF IDEAS

LITERATURE REVIEW	FINDING LITERATURE	FINDING EXAMPLES FROM KNOWN LITERATURE OR ADDING LITERATURE FOR EXISTING STATEMENTS	COMPARING LITERATURE

METHODOLOGY	PROPOSING NEW SOLUTIONS TO PROBLEMS	FINDING ITERATIVE OPTIMIZATIONS	COMPARING RELATED SOLUTIONS

EXPERIMENTS	DESIGNING NEW EXPERIMENTS	EDITING EXISTING EXPERIMENTS	FINDING, COMPARING, AND AGGREGATING RESULTS

WRITING	GENERATING NEW TEXT BASED ON INSTRUCTIONS	ASSISTING IN IMPROVING OWN CONTENT OR PARAPHRASING RELATED WORK	PUTTING OTHER WORKS IN PERSPECTIVE

PRESENTATION	GENERATING NEW ARTIFACTS	IMPROVING THE AESTHETICS OF ARTIFACTS	FINDING RELATIONS BETWEEN OWN OR RELATED ARTIFACTS
		Github Copilot	

CODING	GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE	REFACTORING AND OPTIMIZING EXISTING CODE	COMPARING ASPECTS OF EXISTING CODE

DATA	SUGGESTING NEW SOURCES FOR DATA COLLECTION	CLEANING, NORMALIZING, OR STANDARDIZING DATA	FINDING RELATIONS BETWEEN DATA AND COLLECTION METHODS

ETHICS	WHY DID WE USE AI FOR THIS PROJECT?	WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI?	WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI?
	Consistency	Only used for LaTeX formatting	Only used for LaTeX formatting
	Efficiency / Speed		
	Expertise Access		