University of Applied Sciences and Arts
of Southern Switzerland
SUPSI

Digital History Switzerland, 12.09.2024

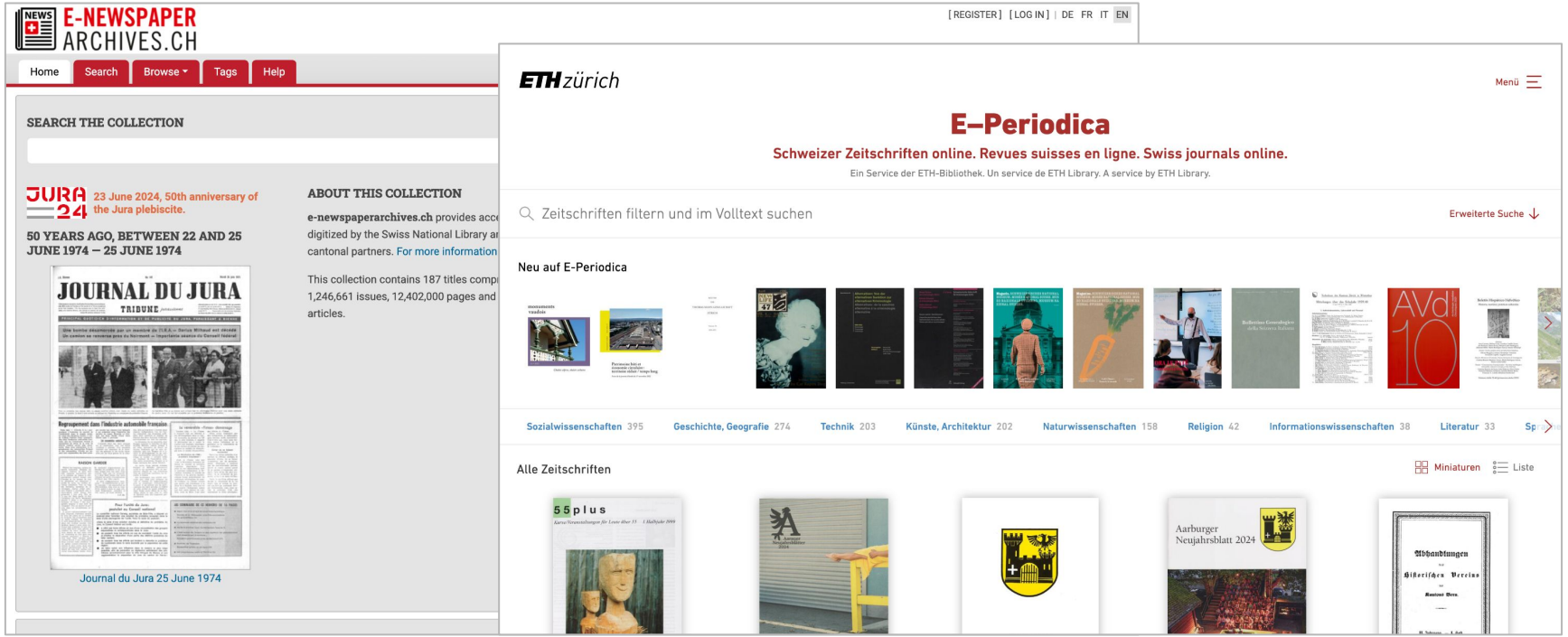
AI-assisted search for digitized publication archives

Fostering the study of historical figures through the use of Natural Language Processing (NLP) and data visualization techniques

Giovanni Profeta, PhD
University of Applied Sciences and Arts of Southern Switzerland (SUPSI)

Context

Archives containing digitized publications

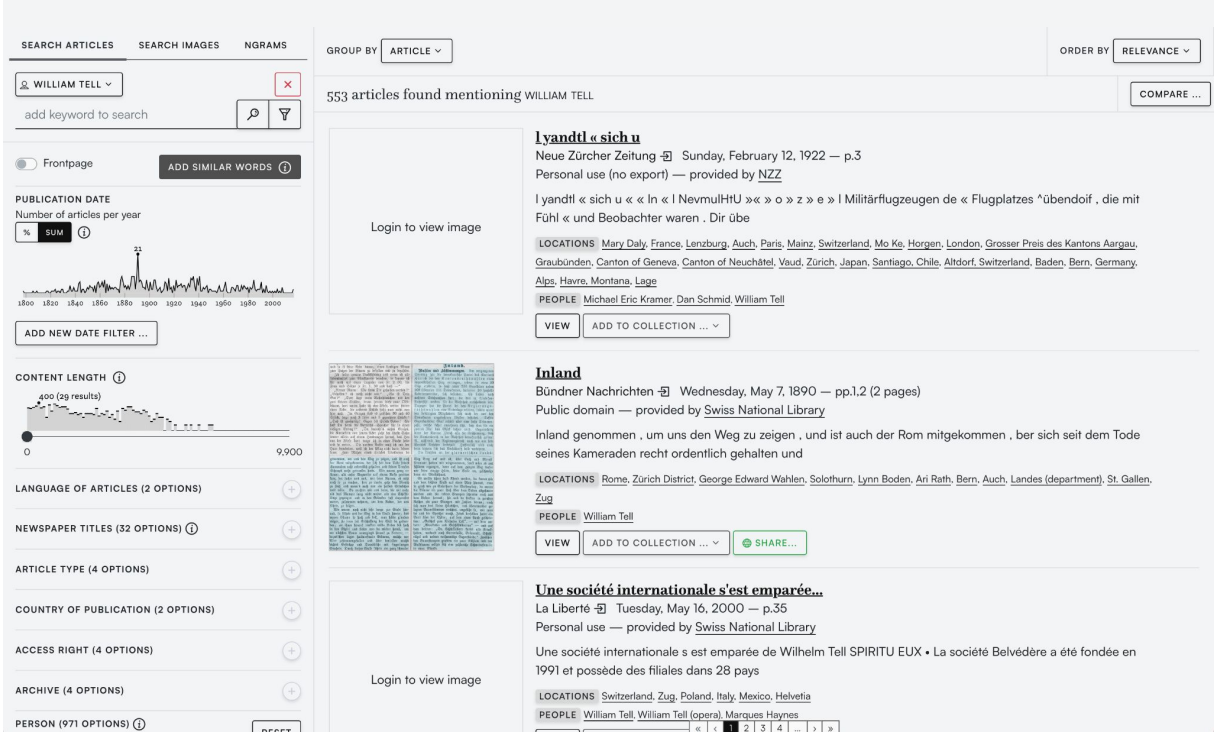


Interface issues in accessing content

- Lack of multiple modalities to investigate an archive
- Lack of explanation on how search results are collected
- Too many results to analyse and lack of advanced search and filtering options
  (it is requested fewer search results but more precision)

Studies on accessing cultural content | Natural Language Process (NLP)

Entity recognition to identify and link related entities

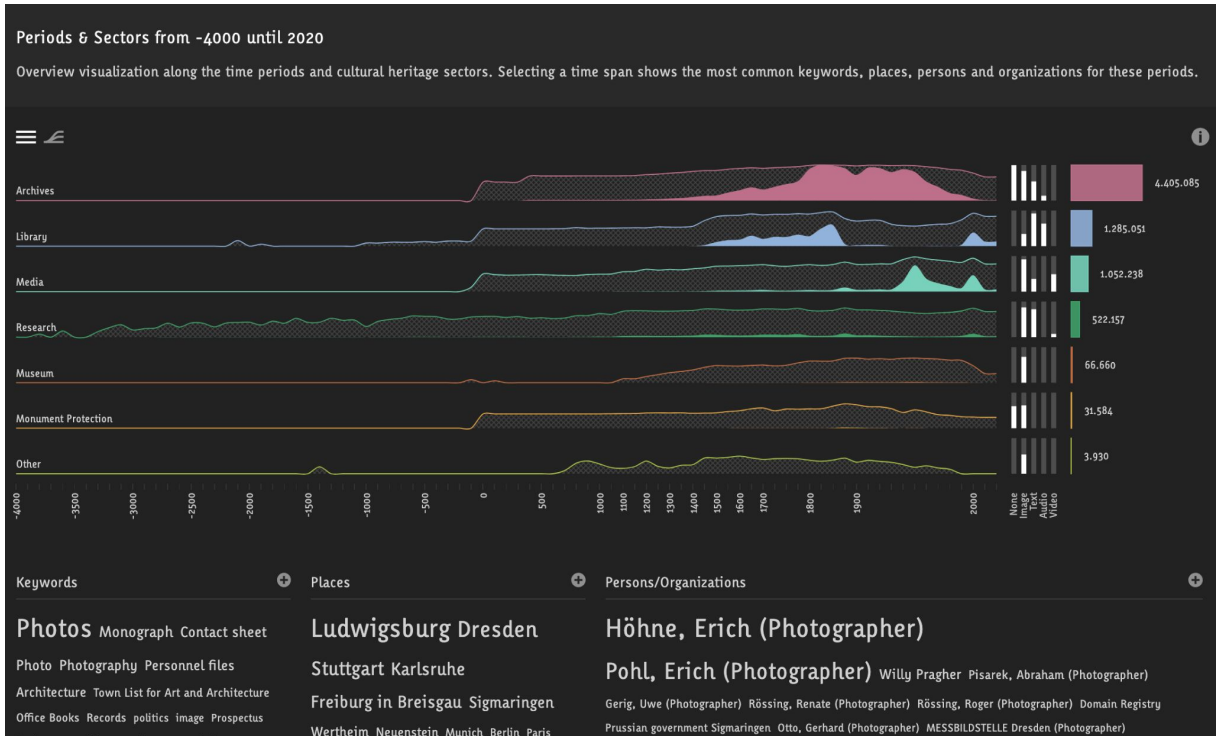


Impresso app, by EPFL, Zurich University, Lausanne University (2017—2020)

Düring, M., Bunout, E., Guido, D.. (2024) “Transparent Generosity. Introducing the Impresso Interface for the Exploration of Semantically Enriched Historical Newspapers”. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(1).

Studies on accessing cultural content | Data visualization

Generous interfaces: multiple metadata visualization as a tool to access cultural content



German Digital Library Visualized, by Urban Complexity Lab (2014)

Whitelaw, M. (2015). “Generous Interfaces for Digital Cultural Collections”. *Digital Humanities Quarterly*, 9(1).

Dörk, M., Pietsch, C., Credico, C. (2017). “One view is not enough: High-level visualizations of a large cultural collection”. *Information Design Journal*, 23(1).

Mini-Muse is a preliminary transdisciplinary project

Main goal

Acquire preliminary knowledge on the integration of AI algorithms and data visualization methods to access and analyse digital libraries.

Research team
Dr. Giovanni Profeta
Dr. Fabio Rinaldi
Joseph Cornelius

Research partner
ETH-Zurich Library
Regina Wanger
Michael Gasser
Christiane Sibille
Michael Ehrismann

The project is funded by Hashler Foundation.

Work plan

WP1	User research and content acquisition
WP2	Development of the NLP algorithms
WP3	Design and implementation of a basic prototype
WP4	Evaluation of the prototype

WP1: User research and content acquisition

Goals

To gather the needs of users of cultural digital archives in terms of types of information needed for their researches

Activities

- Conduction of semi-structured user interviews
- Archival content converted in a format suitable for the NLP algorithms

Deliverables

- Set of user desiderata
- Archival content for the NLP algorithms

Pool of experts involved in the user research

14 heavy users of cultural digital archives interviewed (6 experts in digital history, 8 not experts)

Profession	n	%
historian	6	42.8
historian student	2	14.3
librarian	1	7.1
documentalist	1	7.1
journalist	1	7.1
other (developer, computer scientist)	3	21.6

Location	n	%
Bern	4	28.6
Basel	3	21.4
Zürich	2	14.3
Lugano	2	14.3
Milan	2	14.3
Lausanne	1	7.1

Age range	n	%
20 - 29	2	14.3
30 - 39	2	14.3
40 - 49	6	42.9
50 - 59	2	14.3
60 - 69	2	14.3

Desiderata

Information and features of interest for the users involved in the study

Desiderata	Focus	Request*
1. Extrinsic elements (about the publication) Author of the publication, publication date, title (and subtitle) of the publication, type of document, collocation in the archive/ID	content	Very high (80%-100%)
2. Intrinsic elements (about the content of the publication) Actor (person, institution, country, etc.), action (what the actor did), location, the time in which the action took place;	content	High (60-79%)
3. Getting the link to the point where the information comes from	feature	High (60-79%)
4. Getting results in the language selected by the user (translated if it is needed)	content	High (60-79%)

* The request column shows the percentage of interviewed users which express that desiderata.

Desiderata

Information and features of interest for the users involved in the study

Desiderata	Focus	Request
5. Advanced search (or filter options) Date, and timespan (even a very short timespan)	feature	Medium (40-59%)
6. Need for algorithm transparency Methodology and technology adopted to elaborate a filter/prompt and return an output	(meta) content	Medium (40-59%)
7. Search or filter content in a specific language	feature	Low (20-39%)
8. Action flow: what happens from the moment A to the moment Z	content	Low (20-39%)
9. To compare different points of view (of authors/journals) on the same topic or subject	feature	Low (20-39%)
10. Getting information about the rights of use	content	Low (20-39%)
11. Getting links to other resources for expanding/improving the research (LOD principles)	content	Low (20-39%)
12. See similar articles (similar by topic)	content	Low (20-39%)
13. Index (or visual overview) of the whole collection (titles and authors)	content	Low (20-39%)

Content preparation

1.
Swiss History Journal content, coming from E-Periodica, converted in a format suitable for the NLP algorithms

2.
Selection of 25 recent articles with the following attributes:

- about politics
- in German
- more than one article coming from the same issue
- with shared historical figures

Preliminary word cloud of historical figures

WP2 and WP3: Development of a basic prototype

Goals

To test the feasibility of users' desiderata

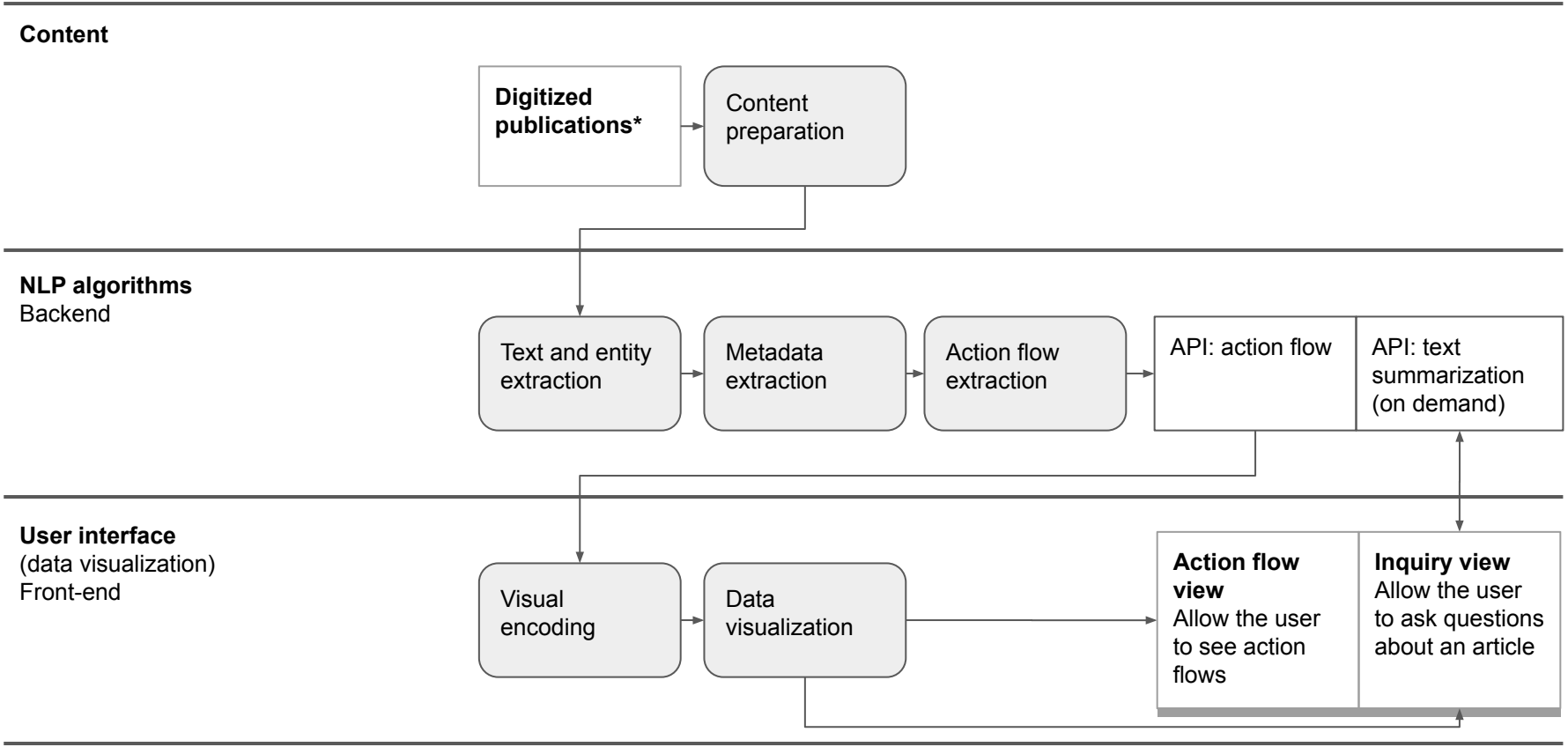
Activities

- Implementation of a set of NLP algorithms
- Implementation of a backend API
- Design and implementation of the frontend

Deliverables

- A basic interactive prototype containing a set of user interfaces of the archive

Mini-Muse information system

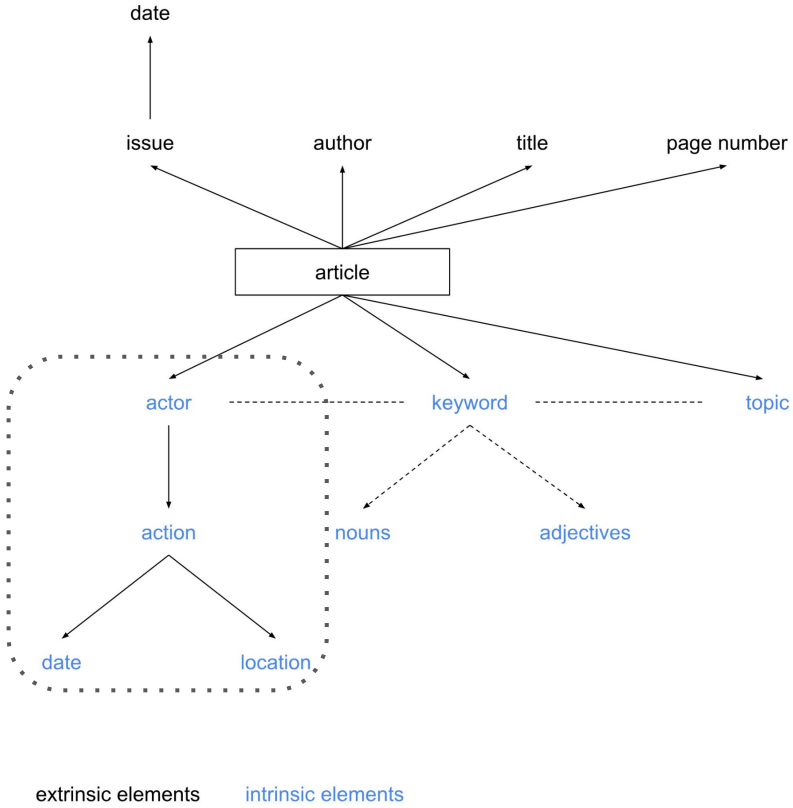


* Digitized publications as XML and “E-Periodica” files.

Text and entity extraction

1.
Selection of a set of type of actions
(active, about doing something)

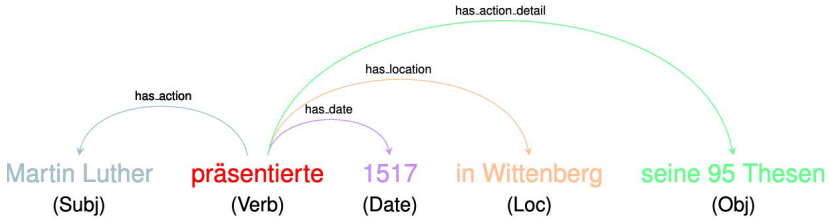
2.
Definition of a set of annotation guidelines
(types of entity to be detected)



Network of entities

NLP algorithms
Action flow extraction

1.
Integration of rule-based algorithms (Text Parsing, NER, Dependency Parsing, and Rule-Based Systems)
Extracting a set of elements: action, actor, object, details, time, and location.



2.
Integration of LLM-based algorithms
Leveraging models like GPT-4 that use structured prompting and large text inputs to understand and interpret complex contexts, identifying actions, actors, and relationships, and inferring details like time and location for accurate action flow detection.

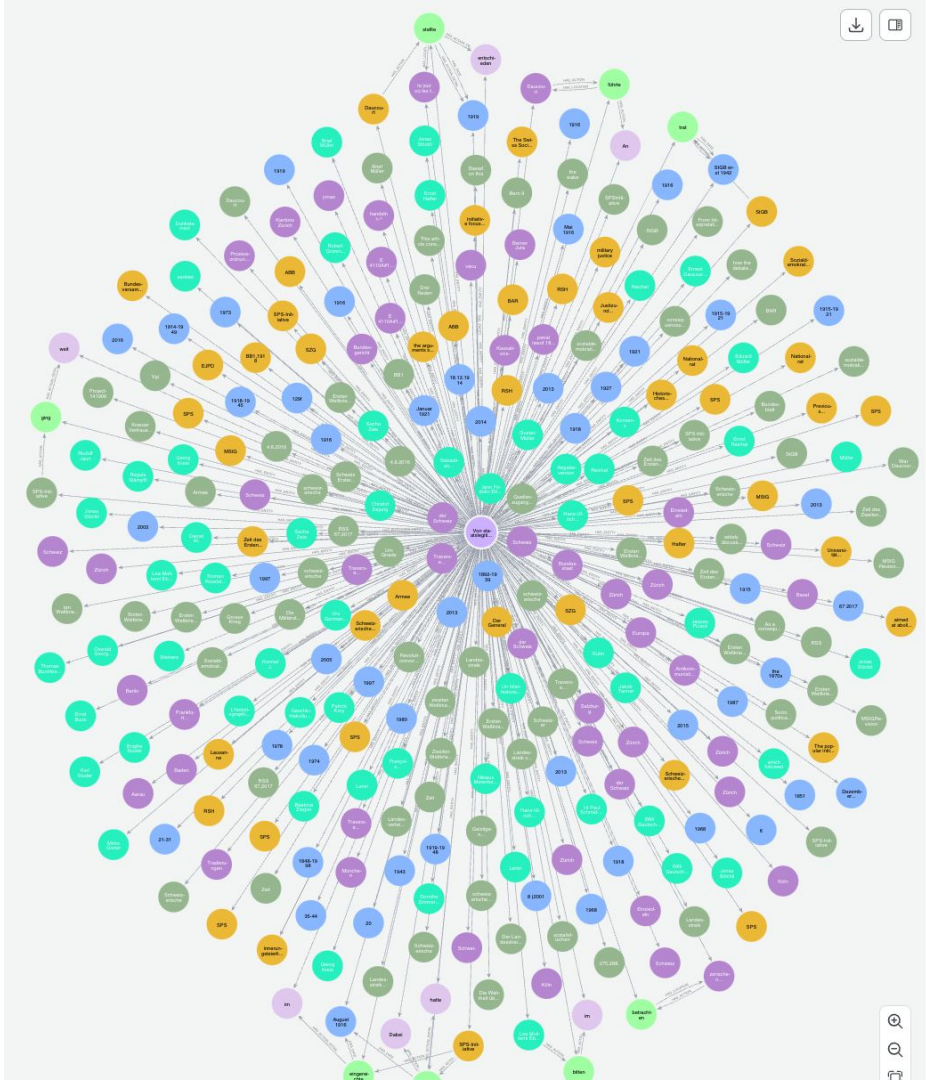
```
Format: JSON
Analyze the following text and extract the action flow, identifying the
actions, actors, objects, details, time, and location:

Martin Luther schlug 1517 seine 95 Thesen an die Tür der Schlosskirche in
Wittenberg. Diese Tat stellte die Praktiken der katholischen Kirche in
Frage, insbesondere den Ablasshandel. Seine Handlungen lösten eine
religiöse Bewegung aus, die als Reformation bekannt wurde. Die 95 Thesen
wurden schnell in ganz Europa verbreitet und verstärkten Luthers Ruf nach
Reformen.
```

```
{
  "Aktion": "stellte in Frage",
  "Akteur": "Martin Luthers Thesenanschlag",
  "Objekt": "die Praktiken der katholischen Kirche",
  "Details": "löste die Reformation aus und verbreitete sich schnell in ganz
  "Zeit": "1517",
  "Ort": "Wittenberger Schlosskirche"
}
```

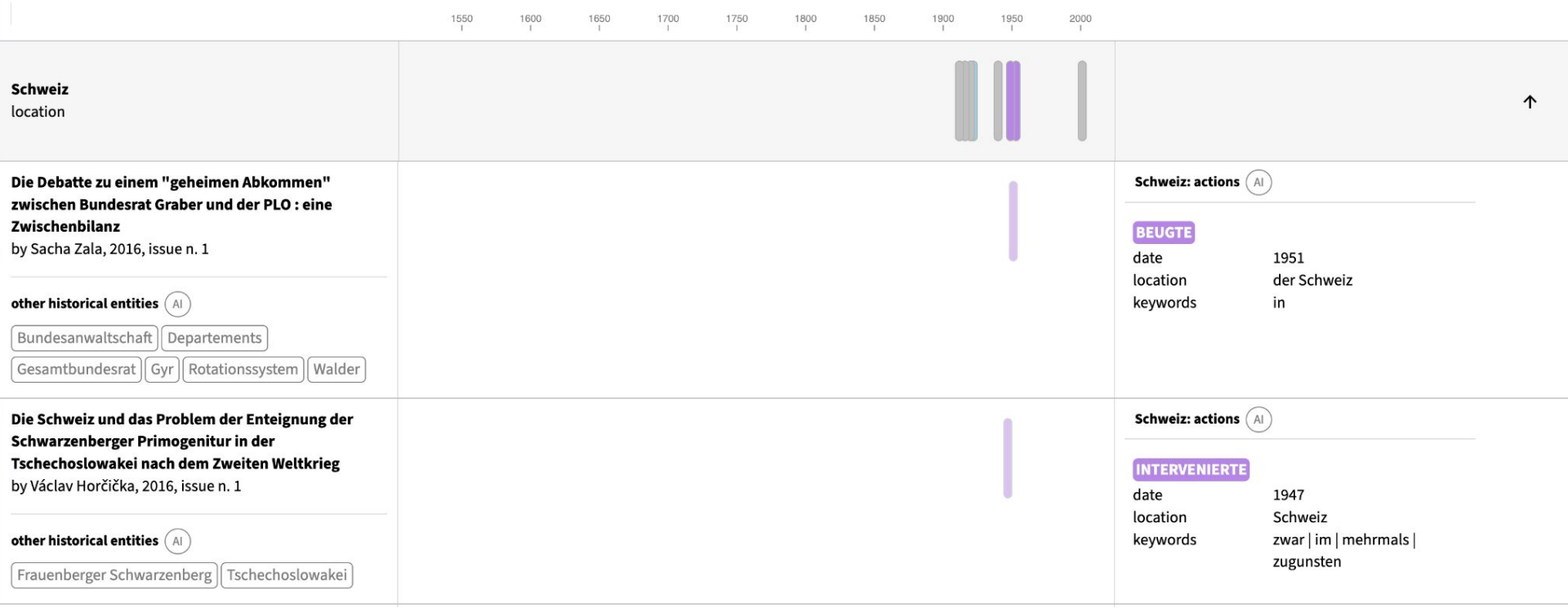
Backend API

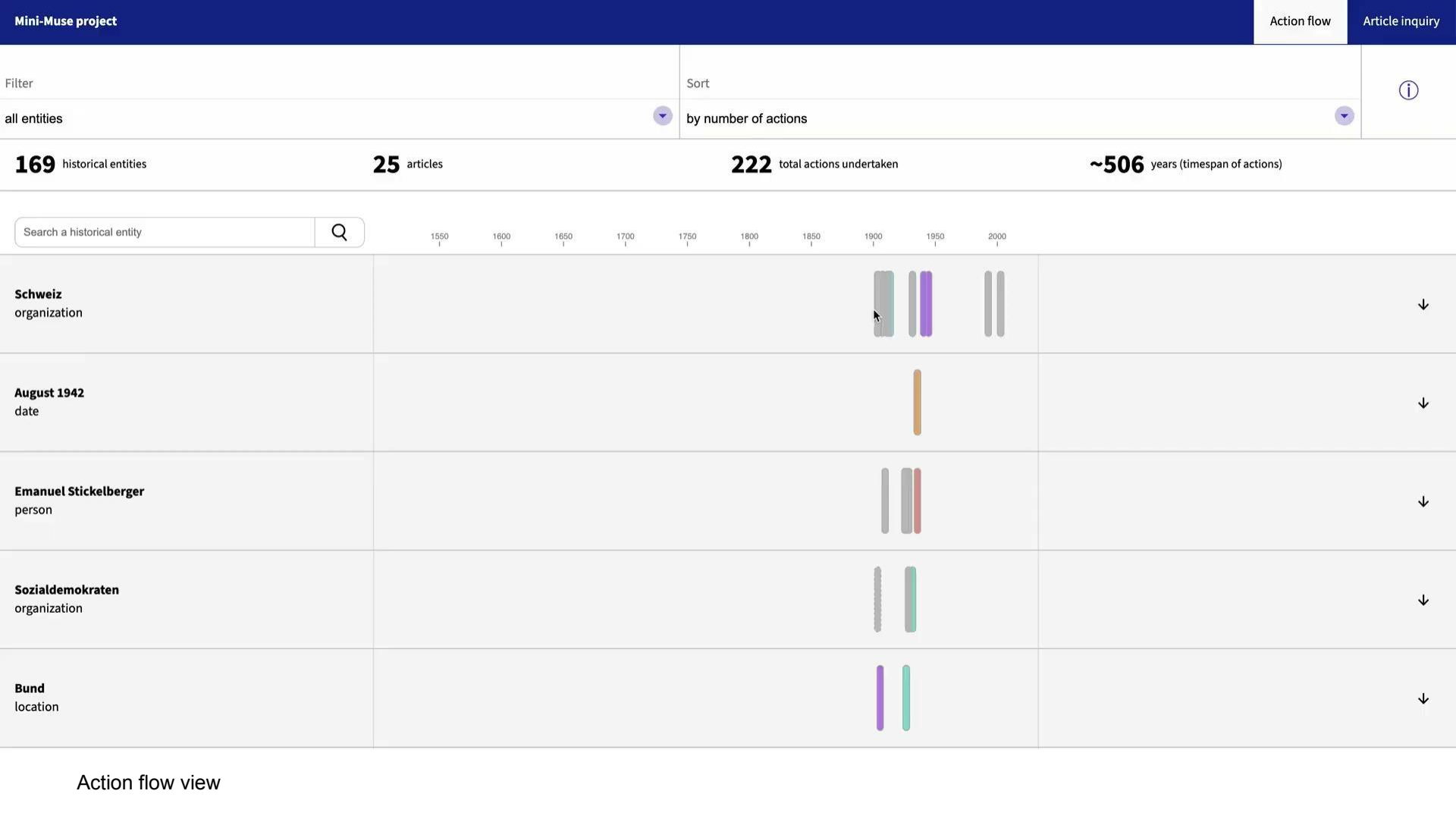
1.
Efficiently store annotations, extracted content, and summarized documents within a graph database to enable data retrieval, and relationship mapping.

2.
Ensure easy accessibility and integration of various annotations and services through a secure web API.



Data visualization





Action flow view

Sort

by publication date

169 historical entities | **25** articles | **222** total actions undertaken | **~506** years (timespan of actions)

Search an author

1600 1700 1800 1900 2000

Der 4. August 1942 : Entscheidungsakteure der Schweizer Flüchtlingspolitik im Kriegsbundesrat
by Thomas Zaugg, 2023, issue n. 1

Als die Apologetinnen der Männerrepublik die Gretchenfrage der Demokratie stellten : zur Politik der Gegnerinnen des Frauenstimmrechts in der Schweiz, 1919-1971
by Noemi Crain Merz, 2022, issue n. 1

"Ausgestreckte Fühler deutscher Gelehrter" : die Universität Basel und akademische Flüchtlinge in den 1930er-Jahren
by Stefanie Mahrer, 2022, issue n. 1

Die interkantonale Dimension der Wohlfahrt : Subsidiarität und die Finanzierung des Heimsektors für Minderjährige
by Alan Canonica, 2022, issue n. 2

Als die Apologetinnen der Männerrepublik die Gretchenfrage der Demokratie stellten : zur Politik der Gegnerinnen des Frauenstimmrechts in der Schweiz, 1919-1971

by Noemi Crain Merz, 2022, issue n. 1
volume n. 72, pp. 40-54

abstract AI

Auch heute dominiert das Bild von Frauen, die sich vor allem über die soziale Stellung ihrer Ehemänner definierten, in der Literatur zum Frauenstimmrecht, die 50 Jahre nach dessen Einführung nochmals wichtige Ergänzungen erhalten hat. Der zeitliche Bogen reicht von 1919, als die ersten Gegnerinnen sich organisierten und auf sich aufmerksam machten, bis 1971, als in einem neuen politischen und sozialen Kontext ihre letzte Stunde schlug. Jean Jacques Rousseau sowie die noch populäreren helvetischen Nationaldenker Johann Heinrich Pestalozzi und Jeremias Gotthelf nahmen die Frau in den Blick aber nicht als politische Bürgerinnen, sondern als Hausfrauen und Mütter.
SZG/RSH/RSS 72/1 (2022), 40-54, DOI: 10.24894/2296-6013.00096 U6 Noemi Crain Merz, Urs Hafner

Ask a question in German about the selected article ...

Article inquiry view

WP4: Evaluation of the prototype (ongoing)

Goals

To gather feedback regarding ease of use, clarity and usability of the interface features

Activities

- Reviews of the prototype according to the feedback received from ETH-Library team
- Conduction of semi-structured user interviews with people involved in the user research
- Anonymous survey with people involved in the user research

Deliverables

- Set of guidelines

Insights gathered

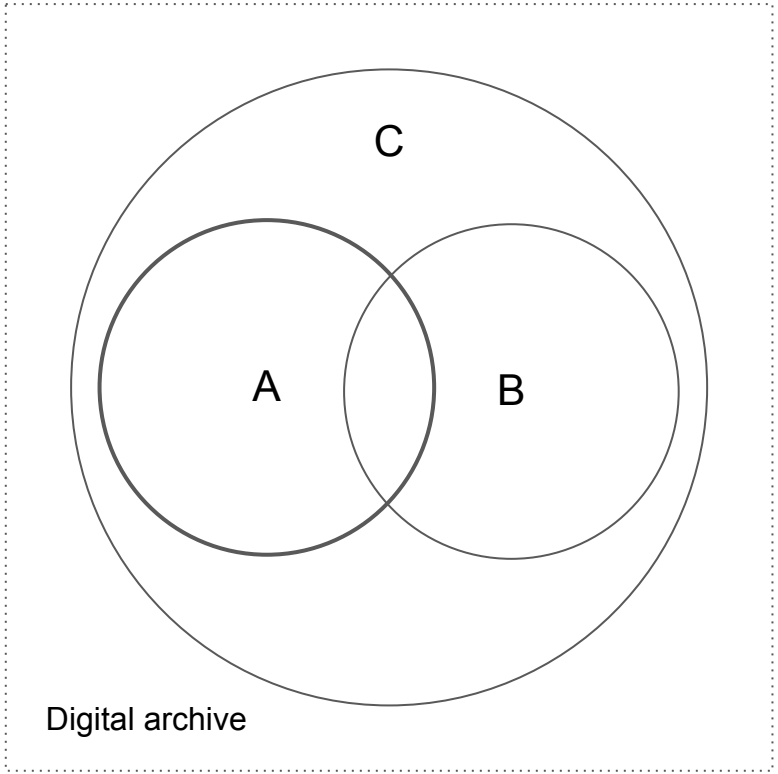
Preliminary results

Aspects of interest

- Showing historical entities' actions and their relationships
- Getting an overview of the action flows
- Getting article summarizations
- Obtaining a clue as to the presence of a certain information

Results

Guidelines for the implementation of cultural digital archives (first draft)



A. Documents centered

B. Historical figures centered

C. Content centered

A. Documents centered

(based on the article inquiry view tests)

Main goal

Allow the user to easily understand which document worth to be read

Provide users with interface features to:

- See correlations among items (figures, authors, dates)
- Summarize the document' content
- Extract meaningful sentences in relations to the user request

B. Historical figures centered

(based on the action flow view tests)

Main goal

Allow the user to easily get the action flows

Provide users with interface features to:

- See figures' action flows
- See correlations among figures (timespan, locations, topics)

C. Content centered
(based on the chatbot tests)

Main goal
Allow the user to easily get information according to its' request

Provide users with interface features to:

- Get an overview of the content
- Summarize content
- Reply questions based on the whole corpus (and provide users with related documents)

Criticalities

- Current OCR systems does not recognize the edges of the articles
- Current size of the prompt of LLMs (not enough for passing very long documents)
- User interface based on web API are very rigid; professionals need a more flexible user interface

Future works

- Investigate how to extract a new historical entity: the thematic area (i.e.: “Bolshevik Revolution”, “Resistance”, “Counter-reform”)
- Investigate how to foster analysis of controversial topics
- Investigate how to foster user contribution (report issues, provide information etc.)

Thank you for your attention.

Project website
https://mini-muse.github.io/project/