Intuitive Semantic Graph Tool for Enhanced Archive Exploration

MARIA TERESA ARTESE AND ISABELLA GAGLIARDI

IMATI - CNR (NATIONAL RESEARCH COUNCIL), MILAN, ITALY

{TERESA,ISABELLA}@MI.IMATI.CNR.IT

The work

To define a pipeline able to create semantic graphs in an Intangible Cultural heritage archive

with two aims:

1. To offer a global view of the content of the archives for any users;
2. To enhance traditional ways of searching and browsing data on the web.

Innovative way to visualize the contents of an archive, as a multilevel graph

Characteristics of the approach

- The pipeline is completely **unsupervised**
- Semantic graphs are **layered**, in case of very large archives
- Use of **pretrained language** models makes it suitable also for a few hundred items
- **Multilanguage** o language specific pretrained language models make it suitable for documents in English, Italian, French, even mixed.

Innovative elements of pipeline

1. **UMAP** is a dimensionality reduction technique. UMAP uses manifold learning for mapping high-dimensional data to a lower-dimensional space while preserving the local structure of the data

2. **HDBSCAN** is a clustering algorithm that groups similar data points together. HDBSCAN does not require the user to specify the number of clusters or the size of the neighbourhood to be searched. It automatically detects the number of clusters and the shape of clusters

3. **Transformers** are a type of neural network used in natural language processing. Transformers use self-attention mechanisms to focus on different parts of the input and capture long-term dependencies.

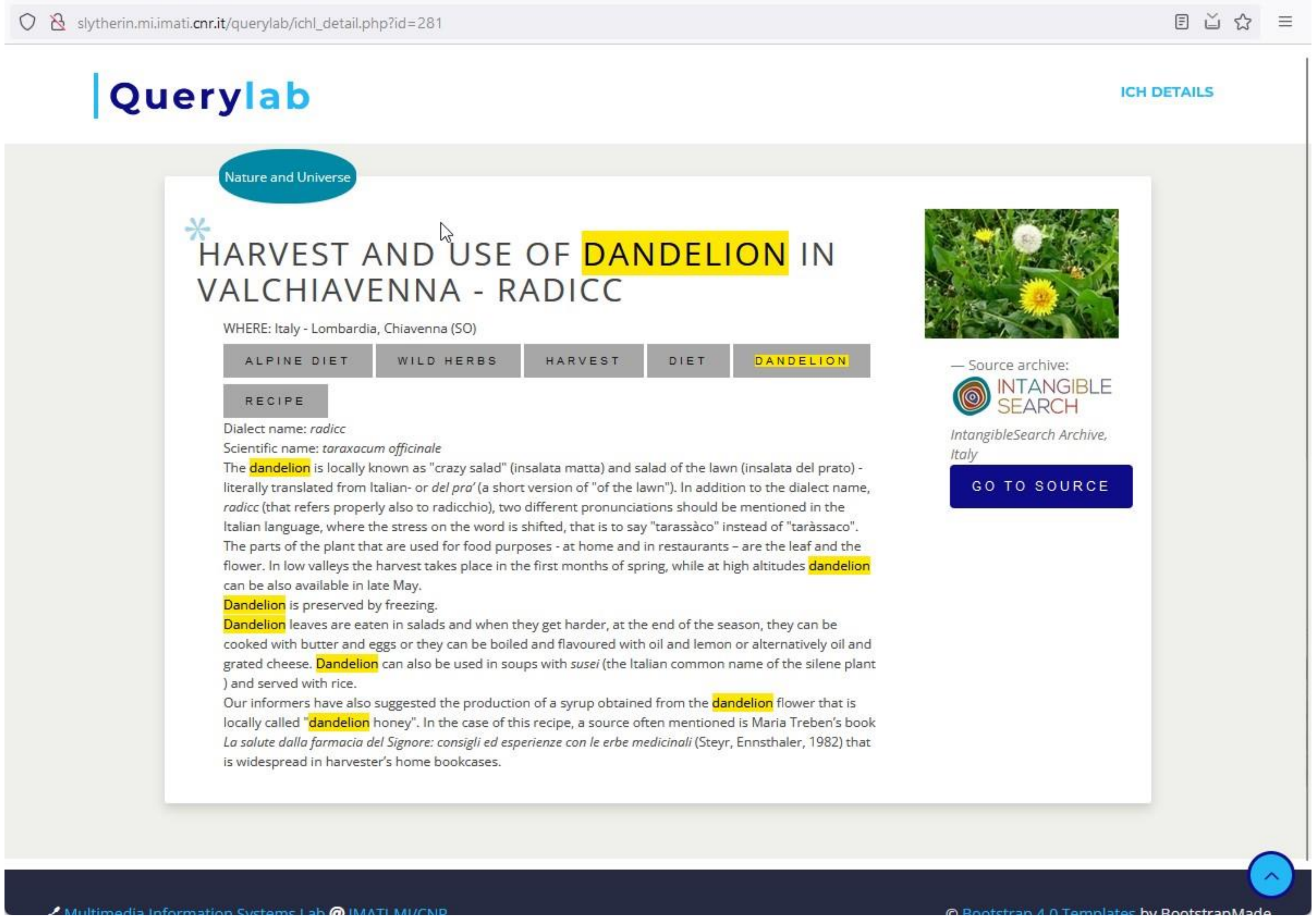
The data: Querylab

A portal specifically designed to manage intangible cultural heritage data.

Two types of data:

- Data stored locally
- and data queried on the fly from remote repositories via REST API web services.

The data: Querylab



HARVEST AND USE OF DANDELION IN VALCHIAVENNA - RADICC

WHERE: Italy - Lombardia, Chiavenna (SO)

ALPINE DIET | WILD HERBS | HARVEST | DIET | DANDELION | RECIPE

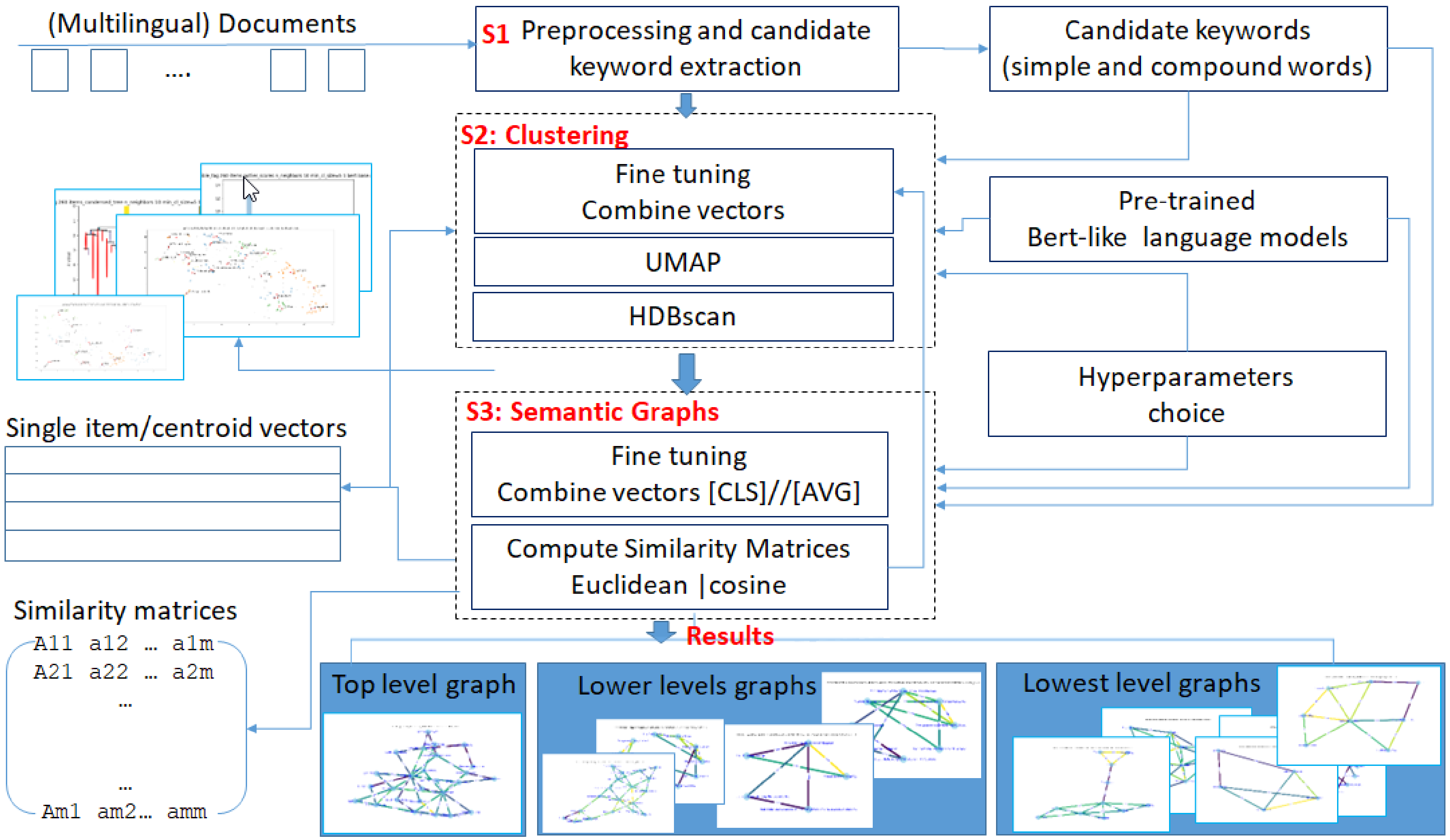
Dialect name: *radicc*
Scientific name: *taraxacum officinale*
The dandelion is locally known as "crazy salad" (insalata matta) and salad of the lawn (insalata del prato) - literally translated from Italian- or *del pra'* (a short version of "of the lawn"). In addition to the dialect name, *radicc* (that refers properly also to radicchio), two different pronunciations should be mentioned in the Italian language, where the stress on the word is shifted, that is to say "tarassàco" instead of "taràssaco". The parts of the plant that are used for food purposes - at home and in restaurants – are the leaf and the flower. In low valleys the harvest takes place in the first months of spring, while at high altitudes dandelion can be also available in late May.
Dandelion is preserved by freezing.
Dandelion leaves are eaten in salads and when they get harder, at the end of the season, they can be cooked with butter and eggs or they can be boiled and flavoured with oil and lemon or alternatively oil and grated cheese. Dandelion can also be used in soups with *susei* (the Italian common name of the silene plant) and served with rice.
Our informers have also suggested the production of a syrup obtained from the dandelion flower that is locally called "dandelion honey". In the case of this recipe, a source often mentioned is Maria Treben's book *La salute dalla farmacia del Signore: consigli ed esperienze con le erbe medicinali* (Steyr, Ennsthaler, 1982) that is widespread in harvester's home bookcases.

— Source archive:
INTANGIBLE SEARCH
IntangibleSearch Archive, Italy

GO TO SOURCE
Multimedia Information Systems Lab @ IMATI MI/CNR
© Bootstrap 4.0 Templates by BootstrapMade

The data: Querylab - 2

1. **Tags**: Expert-defined tags associated with the records in the archive
2. **Title**
3. **Description**
4. **Rake/Textrank Keywords**: simple or compound words were in an automatic and unsupervised manner extracted from descriptions



(Multilingual) Documents

....

S1 Preprocessing and candidate keyword extraction

Candidate keywords (simple and compound words)

S2: Clustering

Fine tuning
Combine vectors

UMAP

HDBscan

Pre-trained
Bert-like language models

Hyperparameters choice

S3: Semantic Graphs

Fine tuning
Combine vectors [CLS]//[AVG]

Compute Similarity Matrices
Euclidean |cosine

Single item/centroid vectors

Similarity matrices

```
A11  a12  …  a1m
A21  a22  …  a2m
          …

          …
 Am1  am2…  amm
```

Results

Top level graph

Lower levels graphs

Lowest level graphs

The proposed pipeline

The proposed pipeline

# Task 1: Dataset Preparation

- Preprocessing (possibly strip stopwords, accents, …)
- Process data to extract items to be used
- Output: items of interest

# Task 2: Items clustering

- Choice of transformers and pre-trained models
- Fine tuning of pre-trained Bert-like models to obtain the vectors
- Choice of hyperparameters for UMAP and HDBSCAN
- Output: centroids of clustered items, and elements of each cluster

# Task 3 Semantic graph creation

- Choice of transformers and pre-trained models, both on raw data and on clustered items and fine tuning
- Creation of similarity matrix using [AVG] or [CLS] tokens
- Output: Semantic graphs con k most similar items, with k=1…4
- Preliminary evaluation of the results with domain experts and web users

Dataset preparation

1. preprocess
2. extract items of interest

↓

Items of interest: either short texts or
set of terms, single or compound words

Items clustering

- Performed using **UMAP and HDBSCAN** on the vectors obtained by tokening items of interest.

  - Choice of hyperparameters for UMAP and HDBSCAN

  - n_neighbor (UMAP) : 20,15,10,5.
  - min_cluster_size (HDBSCAN) 15,10,5,
  - min_samples (HDBSCAN) 15,10,5,1

Items clustering – 2

N neighbors	min_cluster_size	min_samples	Number cluster
20	15	5	11
20	15	1	12
20	10	5	14
20	10	1	15
20	5	5	17
20	5	1	30
15	10	1	18
15	5	1	34
10	10	5	13
10	10	1	20
10	5	5	18
10	5	1	33
5	10	1	21
5	5	5	26
5	5	1	45

Items clustering – 3

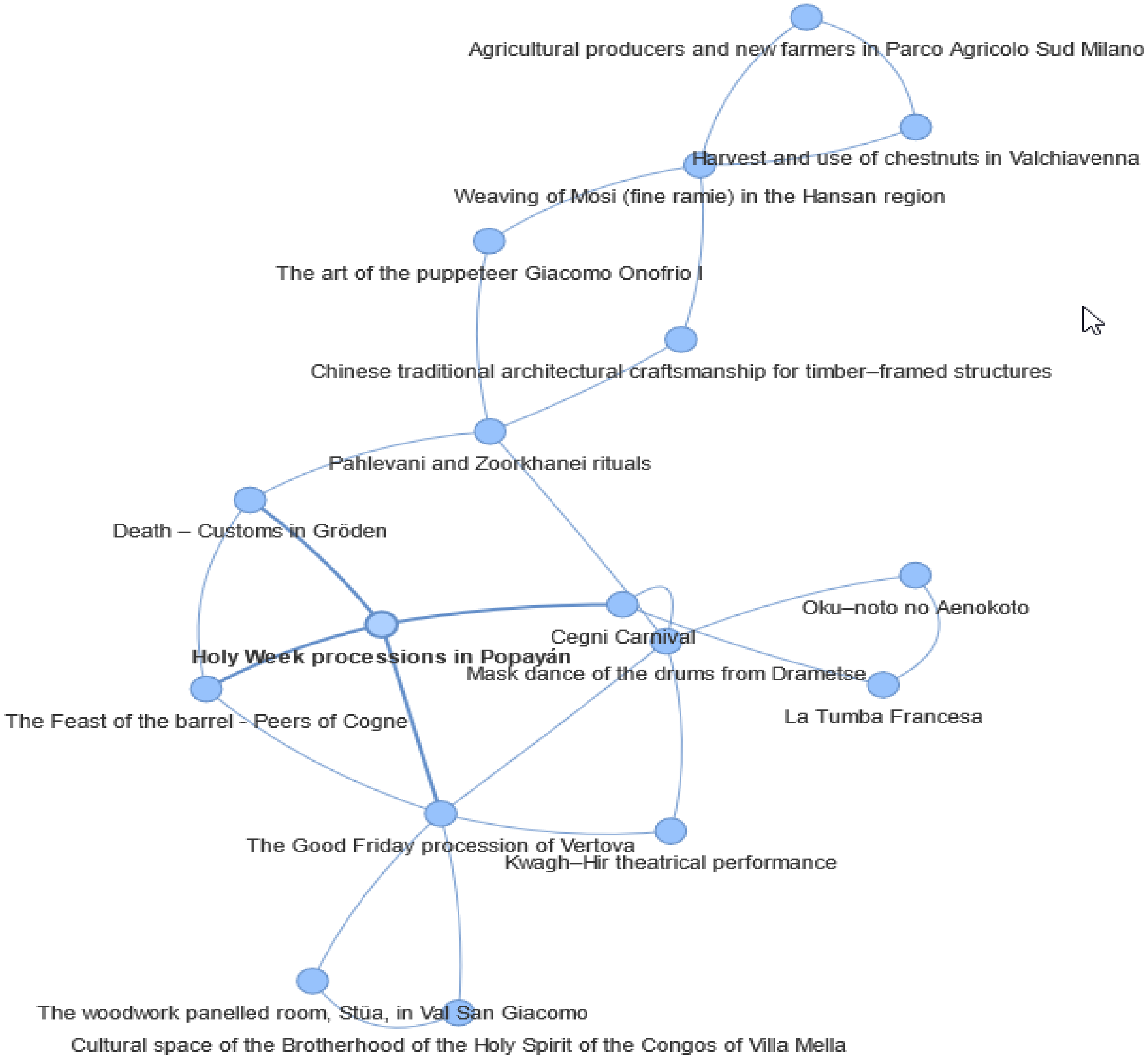
- Performed using UMAP and DBSCAN on the **vectors obtained by tokening items of interest.**
  - Choice of transformers and pre-trained languages among these:
  - **BERT Base**: This is the original pre-trained BERT model released by Google. It has 12 transformer layers and is trained on a large corpus of text data from Wikipedia and the Book Corpus dataset.
  - **DistilBERT**: a distilled version of BERT model: smaller, faster, cheaper and lighter.
  - **MiniLM-L6-v2**: This is a smaller version of the BERT model developed by Microsoft. It has only 6 transformer layers and is trained on a subset of the data used to train BERT Base.
  - **Bert-base-Wikipedia-sections-mean-tokens**: This is a pre-trained BERT model released by the Hugging Face team. It is trained on a large corpus of text data from Wikipedia and uses a mean pooling strategy to create a fixed-length representation of the input text.

Semantic Graph Creation

- Creation of similarity matrices, for the centroids and the clusters
- Use of [CLS] or [AVG] to create a single vector per item (title, tag, description, …)
- K-most similar items, with k ranging from 1 to 4

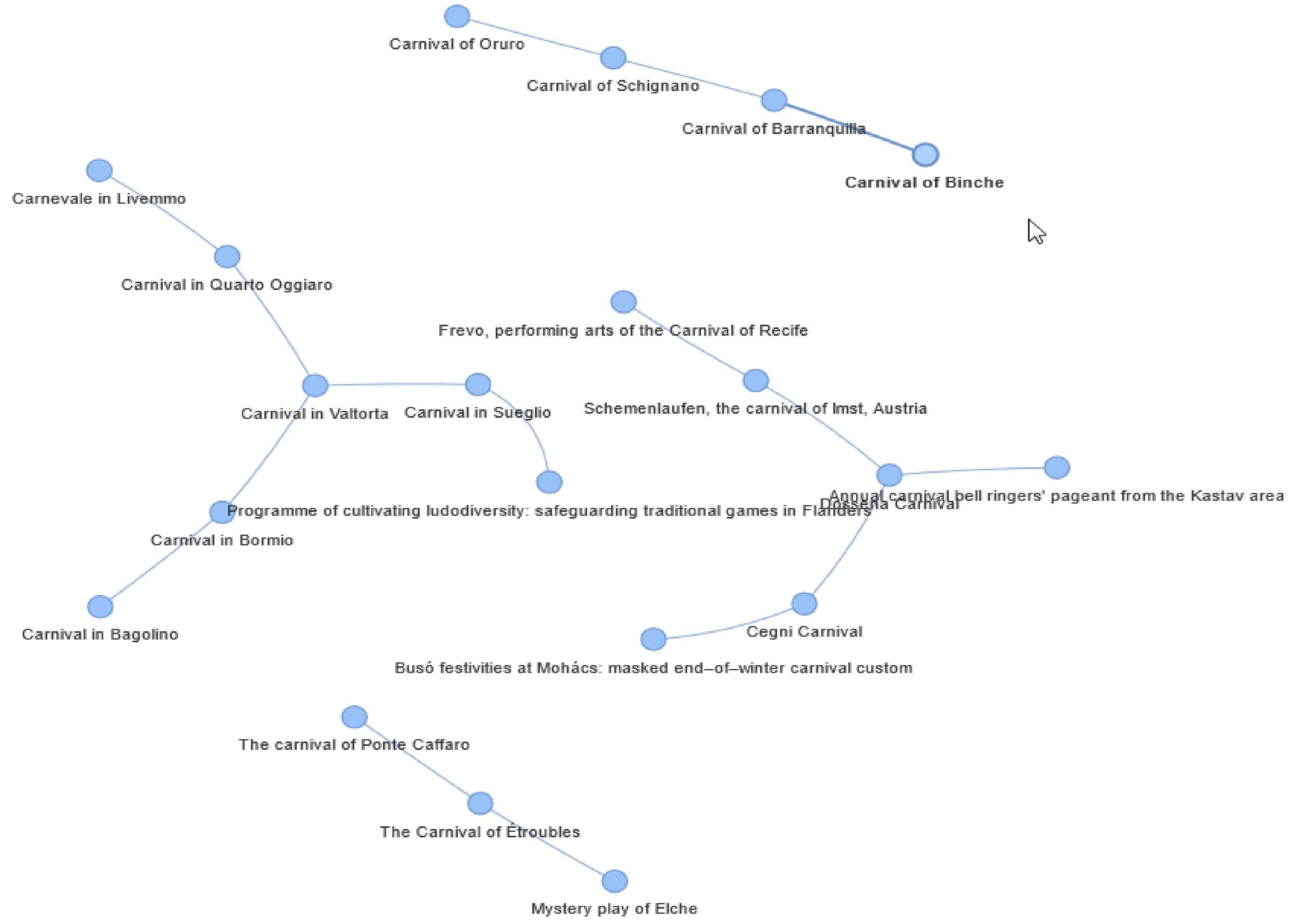
Whole dataset

With k=2


Agricultural producers and new farmers in Parco Agricolo Sud Milano
Harvest and use of chestnuts in Valchiavenna
Weaving of Mosi (fine ramie) in the Hansan region
The art of the puppeteer Giacomo Onofrio I
Chinese traditional architectural craftsmanship for timber–framed structures
Pahlevani and Zoorkhanei rituals
Death – Customs in Gröden
Oku–noto no Aenokoto
Cegni Carnival
Holy Week processions in Popayán
Mask dance of the drums from Drametse
La Tumba Francesa
The Feast of the barrel - Peers of Cogne
The Good Friday procession of Vertova
Kwagh–Hir theatrical performance
The woodwork panelled room, Stüa, in Val San Giacomo
Cultural space of the Brotherhood of the Holy Spirit of the Congos of Villa Mella

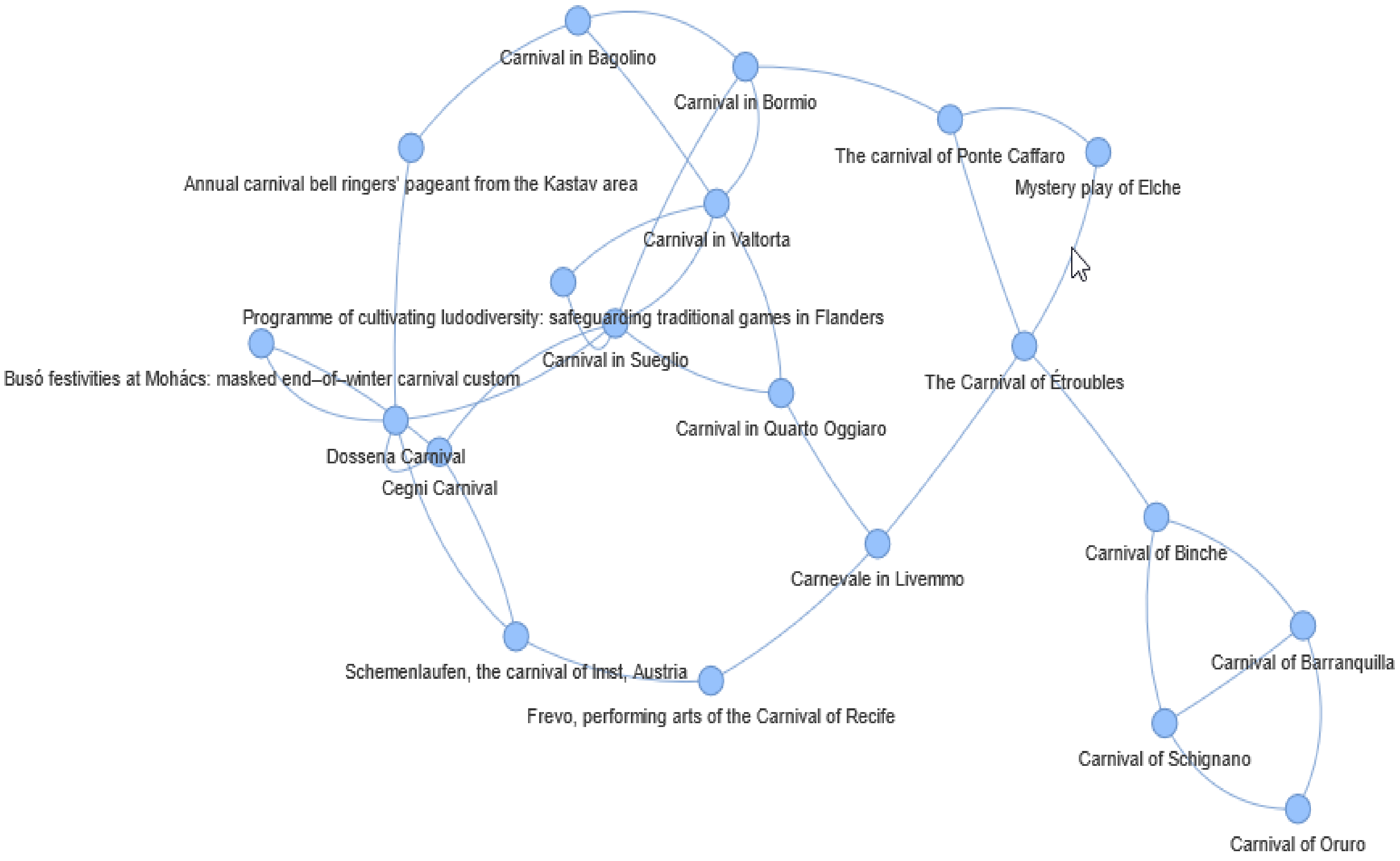
Cegni Carnival

With k=1


Carnival of Oruro
Carnival of Schignano
Carnival of Barranquilla
Carnival of Binche
Carnevale in Livemmo
Carnival in Quarto Oggiaro
Frevo, performing arts of the Carnival of Recife
Carnival in Valtorta
Carnival in Sueglio
Schemenlaufen, the carnival of Imst, Austria
Annual carnival bell ringers' pageant from the Kastav area
Dossena Carnival
Programme of cultivating ludodiversity: safeguarding traditional games in Flanders
Carnival in Bormio
Carnival in Bagolino
Cegni Carnival
Busó festivities at Mohács: masked end–of–winter carnival custom
The carnival of Ponte Caffaro
The Carnival of Étroubles
Mystery play of Elche

Cegni Carnival

With k=2



Carnival in Bagolino
Carnival in Bormio
The carnival of Ponte Caffaro
Mystery play of Elche
Annual carnival bell ringers' pageant from the Kastav area
Carnival in Valtorta
Programme of cultivating ludodiversity: safeguarding traditional games in Flanders
Carnival in Sueglio
Busó festivities at Mohács: masked end–of–winter carnival custom
The Carnival of Étroubles
Carnival in Quarto Oggiaro
Dossena Carnival
Cegni Carnival
Carnival of Binche
Carnevale in Livemmo
Carnival of Barranquilla
Schemenlaufen, the carnival of Imst, Austria
Frevo, performing arts of the Carnival of Recife
Carnival of Schignano
Carnival of Oruro

Evaluation

Qualitative evaluation

We gathered feedback from heritage experts and web users:

- is the clustering and similarity matrix able to extract the significant elements?, and
- Do users find browsing the archive via graph interesting and useful?

- **Positive results:**
  - **PROs**: simplicity and usability of the graph visualization
  - **CONs**: low-level clusters contained elements that were not closely related, or that some related elements were spread across multiple clusters

Conclusion

Definition of a pipeline for the creation of semantic graphs as a layered map with different granularity

New way of searching and browsing ICH archives

Preliminary evaluation

- Effectiveness of the pipeline in generating meaningful semantic graphs
- Positive evaluation from users, but graphs with more than 30 nodes (overly dense graphs) are difficult to understand and navigate

Future works

- tools to traverse the graphs
- fish-eye views to overcome the overly dense graphs
- experiments on other datasets

Thank you!

For any question

Isabella Gagliardi	isabella@mi.imati.cnr.it
Maria Teresa Artese	teresa@mi.imati.cnr.it

www.imati.cnr.it

arm.mi.imati.cnr.it/querylab

arm.mi.imati.cnr.it/mislab

arm.mi.imati.cnr.it/papers/ht2023 for additional materials