

DARIAH Annual Event Lisbon 2024 “Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities”

Paper Proposal

Title: Odeuropa Workflow Use Case: Olfactory Digital Data from the Digital Library of Slovenia

Contributors:

Ines Vodopivec, National and University Library of Slovenia, 0000-0002-3674-8630 (presenting author)

Authors:

Inna Novalija, Jožef Stefan Institute, Slovenia, 0000-0003-2598-0116

Dunja Mladenčić, Jožef Stefan Institute, Slovenia, 0000-0002-0360-6505

Pasquale Lisena, EURECOM, France, 0000-0003-3094-5585

Raphael Troncy, EURECOM, France, 0000-0003-0457-1436

Inger Leemans, Royal Netherlands Academy of Arts and Sciences (KNAW), Vrije Universiteit Amsterdam, 0000-0003-1640-4109

Abstract

The Odeuropa project integrated expertise in sensory mining, knowledge representation, computational linguistics, (art) history, and heritage science. Digital data were extracted from thousands of images and historical texts in six languages, all available in the public domain through the Smell Explorer or the Encyclopaedia of Smell History and Heritage.

Digital textual collections integrated into the workflow are represented in the Knowledge Graph, including eighteen GLAM institutions. Among others, these institutions include the British Library, Digitale Bibliotheek voor de Nederlandse Letteren, Deutsches Text Archiv, Digital Library of Slovenia, Europeana, Gallica, and WikiSource. Cooperating GLAMs collaborated with researchers from the UK, Netherlands, Germany, Italy, France, and Slovenia. In this paper, the results of inter- and transdisciplinary work are presented.

A variety of smell experiences, described by smell words, smell sources, qualities associated with smells, smell perceivers, etc., have been extracted from the available digital data. Books (monographs and dissertations) and periodicals (historical, scientific, general newspapers, and journals) have been the primary resources for Odeuropa smell data analysis. Additionally, manuscripts (mediaeval codices and literary manuscripts); images (photographs, postcards, posters); music (musical scores and audio recordings); and maps (maps and atlases) were used.

The Smell Experiences extraction workflow applied in the Odeuropa project encompassed six main steps: (1) Development of annotated benchmarks, (2) Analysis of benchmark statistics, (3) Development of a text processing system, (4) Extraction of Smell experiences, (5) Linking with Odeuropa semantic vocabularies, and (6) Exploration, visualization, storytelling. The workflow enabled the integration of a wide variety of resources for digital humanities research, from textual materials to visual representations.

Under step one, ten domains of interest were defined, such as Household & Recipes or Perfumes & Fashion. The project aimed to include 10 documents for each category in the benchmark, totalling 100 documents. Next, annotation setting was performed for each language using the INCEpTION tool.¹ Quality control of the annotated datasets, addressing issues such as missing annotations, smell words with double annotations, unlinked frame elements, and relation errors, was then conducted by the researchers.

Analysing benchmark statistics was the second step. From the statistics of the Slovenian annotated benchmark, it became apparent that smell sources and qualities are frequently found together with smell words in historical texts.

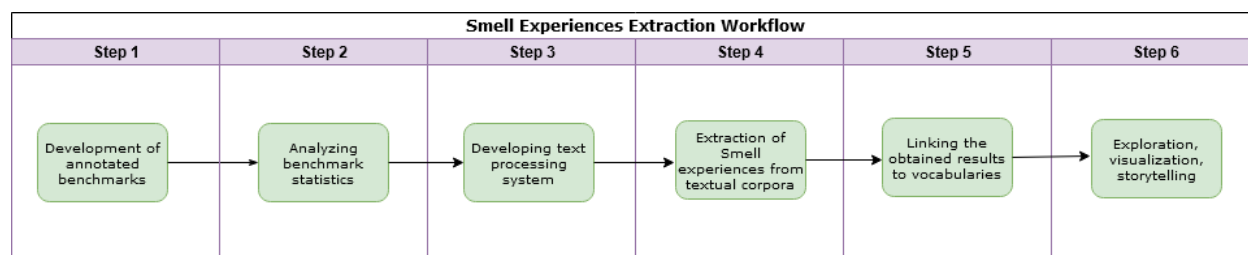
Under step three, the annotated benchmark was utilized to develop models for extracting smell frame elements from the text corpora. One of the main outcomes was the development of state-of-the-art smell extraction models for included languages. Sloberta² served as the basis for developing Slovenian one.

Subsequently, Odeuropa partners utilized the developed models to extract the actual smell data for the European Olfactory Knowledge Graph in step four. This was followed by linking of data with Odeuropa semantic vocabularies³ and embedding them into Odeuropa tools in step five. The Slovenian dataset was digitally processed, smell experiences were obtained and integrated into the Odeuropa Smell Explorer tool.⁴

In the last stage, Odeuropa results were used for exploration, visualization, and storytelling. The number of tools and resources can be reused in the context of digital cultural heritage.

Figures

Figure 1: Smell Extraction Workflow



¹ Klie, J.-C., Bugert, M., Boulosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In Proceedings of the 27th International Conference on Computational Linguistics: System. Demonstrations, pages 5–9. Association for Computational Linguistics.

² Ulčar, M. and Robnik-Šikonja, M. (2021). Sloberta: Slovene monolingual large pretrained masked language model.

³ Odeuropa vocabularies [online]. Available from: <https://vocab.odeuropa.eu> (accessed in February 2024).

⁴ Odeuropa Smell Explorer [online]. Available from: <https://explorer.odeuropa.eu> (accessed in February 2024).

Figure 2: Example of a textual resource in Slovenian language available in Odeuropa Smell Explorer tool (Book of Sirach, The Praise of Wisdom, 1584)

 Besedilni vir

Sveto pismo (Jurij Dalmatin)


 Save
Odeuropa benchmark [\(external link\)](#)

<p style="color: #4a7ebb; font-weight: bold;">Smell Emission</p> <p>DATUM 1584 NASTANKA</p>	<p style="color: #4a7ebb; font-weight: bold;">Olfactory Experience</p> <p>OPREDELEJEN KOT lubesniu JEZIK sl</p>
---	---

 Izvleček 1 ▼

W [DAL Sir 24:15] Ieſt fim en lubesniu duh dala od febe, kakor Zimet, inu kakor enu shlahtnu diſhezhe korenje, inu kakor ta ner bullni Myrra, kakor Galban inu Onix, inu Myrra, inu kakor kadilu v'Templi.

Figure 3: Image Annotations in Odeuropa Smell Explorer, Flower piece, Sam Segal, 1650



Annotations | select all (32)

- Bivalve (6)
- Butterfly (4)
- Caterpillar (1)
- Dragonfly (1)
- Flower (12)
- Forget-me-not (1)
- Insect (1)
- Iris (1)
- Petunia (1)
- Tulip (4)