Art DATIS: Improving Search in Multilingual Corpora to Support Art Historians

Vera Provatorova, Evangelos Kanoulas, Anna Carlgren, Sven Dupré, Marieke Hendriksen
Informatics Institute, University of Amsterdam
P.O. Box 94323, 1090 GH Amsterdam, The Netherlands
v.provatorova@uva.nl

Keywords: multilingual search, art history corpora, linked open data, digital art history, taxonomy induction

Introduction/Purpose

Working with historical corpora can be a challenging task, especially if the data is heterogeneous: multilingual, multimodal, and including historical documents spanning a long period of time. The project Art DATIS addresses this challenge in the field of digital art history. Art DATIS (Digital Art Technical sources for the Netherlands: Integration and improvement of sources on glass for a Sustainable future) is a four-year research project (2018-2022) within the Netherlands Organisation for Scientific Research (NWO) Big Data / Digital Humanities program. The project is a collaboration between University of Utrecht, University of Amsterdam, RKD Netherlands Institute for Art History, the Free Glass Foundation, and Picturae company who has digitized the data and provides technical support.
The purpose of the project is to develop efficient information retrieval algorithms for the digitized archives of Dutch glass artist Sybren Valkema (1916-1996), and connect the archives to existing data collections on free glass production, making the data easily accessible for art historians and glass artists all over the world. Achieving this goal will be important for many art historians, as Valkema played a key role in founding the international free glass movement, and his archives contain valuable information for research in this field.

Methods

The technical side of the project, in its turn, raises promising research questions in the field of information retrieval: working with Valkema's archives and the relevant sources of data brings up several challenges that require novel solutions. Firstly, the historical sources kept in the archives start from as early as the 16th century, so it is important to consider the changes of language over time. Secondly, the data is multilingual: there are sources in Dutch, English, German and other languages, which makes it necessary to develop an approach that takes multilinguality into account. Thirdly, integrating the historical sources on free glass production requires enriching the existing art history vocabularies with new terms. And, last but not the

least, the sources are highly diverse and heterogeneous, which makes it a challenging task to bring them together into a unified database.

The archives of Sybren Valkema were chosen for the project because his work is the topic of the research questions raised by the art history side of Art DATIS. The results of the data science part of the project are topic-independent and can easily be generalised and applied to other collections of data.

The data is currently digitised in the form of images, most of which contain text in typed or handwritten form. To prepare the data for applying the newly developed algorithms, it is necessary to classify the images that contain typed text versus those that contain handwritten text, and apply different text recognition algorithms respectively. After this step of data processing is finished, the search and tagging algorithms for textual data can be developed and applied to the collection.

Expected results

The project started in September 2018 and will be finished in 2022. Its expected results are of two kinds, algorithms and datasets, described below:

A1. A novel taxonomy induction algorithm that operates over multilingual, historical data, to be applied to the glass-related datasets to enrich the AAT thesaurus with new glass-related terms extracted from the archives.

A2. An automated tagging system, which classifies multilingual, historical corpora against an ontology at different granularities of text (document, paragraph, sentence, term) applied to the glass-related data.

A3. Efficient search algorithms, aware of multilinguality and historical changes of language, that account for both lexical, latent, and semantic relations between text and searchers needs, to be applied to the glass related data, and glass related research.

D1. The archives integrated with existing sources by using linked open data techniques, and structured with an ontology that adheres to international standards for cultural heritage.

D2. The integrated big dataset to be annotated, fully searchable and available via the RKD Explore interface.

References

1. Art DATIS. (2018). Project Art DATIS. [online] Available at: https://artdatis.nl/ [Accessed 30 Jan. 2019].
2. Getty.edu. (2019). Art & Architecture Thesaurus (Getty Research Institute). [online] Available at: http://www.getty.edu/research/tools/vocabularies/aat/ [Accessed 31 Jan. 2019].

3. Rkd.nl. (2019). Explore. [online] Available at: https://rkd.nl/nl/collecties/explore [Accessed 31 Jan. 2019].
4. Wang, C., He, X., & Zhou, A. (2017). A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1190-1203).
5. Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2145-2158).
6. Kenter, T., Wevers, M., Huijnen, P., & de Rijke, M. (2015, October). Ad hoc monitoring of vocabulary shifts over time. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1191-1200). ACM.
7. Steinberger, R. (2012). A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation, 46(2), 155-176.
8. Eide, Ø., Felicetti, A., Ore, C. E., D'Andrea, A., & Holmen, J. (2008, February). Encoding cultural heritage information for the semantic web. procedures for data integration through cidoc-crm mapping. In Open Digital Cultural Heritage Systems Conference (p. 47).
9. Wood, D., Zaidman, M., Ruth, L., & Hausenblas, M. (2014). Linked Data. Manning Publications Co..
10. Baca, M. (Ed.). (2008). Introduction to metadata. Getty Publications.