



Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices

Dominik Gerstorfer <dominik_dot_gerstorfer_at_tu-darmstadt_dot_de>, Technische Universität Darmstadt 

Evelyn Gius <evelyn_dot_gius_at_tu-darmstadt_dot_de >, Technische Universität Darmstadt 

Janina Jacke <janina_dot_jacke_at_uni-goettingen_dot_de>, Georg-August-Universität Göttingen 

Abstract

This is the editorial of the sepecial issue “Working on and with Categories for Text Analysis.”

Why we need new theoretical and methodological perspectives on categories in the digital humanities – a “rolling editorial”

In the realm of digital humanities, computational social sciences and related fields, categories are omnipresent. Not only do we use them to systematize our objects of interest, such as texts or aesthetic artifacts, and organize their representations in databases and repositories. Categories also play an important role in text analysis, especially when extracting or annotating parts of texts in a structured manner. 1

Categories serve as powerful tools for both purposes. They allow, for example, the linguistic labeling of objects or elements and their subsequent grouping according to selected relevant features. In this context, categories seem suited to integrate two important complementary tasks: if they are based on adequate parameters, they can further a sensible *reduction of complexity*, thereby facilitating the analysis of and communication about complex textual artifacts and data sets. At the same time, categories offer the possibility of a *detailed description* through the creation of subcategories. When categories are organized in systems such as ontologies or taxonomies, they can additionally provide information about the relationship between relevant phenomena. Moreover, since creating categories usually requires defining terms explicitly, categories greatly facilitate the scholarly exchange of information on subjects in the humanities, cultural studies, or social sciences – among other things, through enhancing understanding and comparability of claims and hypotheses. 2

The prevalence of categories in the digital humanities can be attributed, in part, to the influence of standards from the formal sciences in this field. In contrast to this, developing and using categories is rather the exception than a rule when we look at most traditional humanities disciplines, where it is limited to certain sub-disciplines. This, together with the omnipresence of categories and at the same time little systematic reflection in the digital humanities, raises a number of questions: Which categories or which types of category systems are appropriate for objects in the humanities? What determines the validity and fruitfulness of categories in this field? How can we develop and revise category systems using existing or new procedures? And how can we employ categories to address complex, and often hermeneutical, questions that are central to most humanities disciplines? 3

Answering these questions requires considering multiple perspectives, not only from different humanities disciplines and social sciences, but also from information science and technology. Taking this as impetus, we have organized two interdisciplinary workshops on the topic of categories in the digital humanities.^[1] The first workshop focused on theoretical and formalistic aspects, exploring non-hierarchical concept ontologies and markup schemas. The second 4

workshop emphasized methodological and application-oriented aspects, specifically the development and application of category systems for text research. Drawing from the inspiring ideas and projects presented and discussed there, we decided to collect the ideas in a special issue. With an open call for papers, we therefore invited the workshop contributors as well as other interested researchers to contribute to this issue.

One focus of this issue lies on the work on and with categories for text annotation and analysis. A second focus is on systems and methods for the organization and classification of texts in the context of databases – as well as the interplay between these two areas where category systems play a crucial role. We requested that the contributions be based on concrete studies in the field of the digital humanities or related fields, or provide an information science perspective that has been, or can be, adapted in the digital humanities.

The contributions that were selected for this issue cover a wide range of work on and with categories. They delve not only the development and application of category systems themselves and the disciplines that inform and require them but also the various conceptual and pragmatic problems encountered during these processes.

This special issue will be published as a rolling issue, with the first bundle of articles in July 2023. Leveraging the advantages of the digital format, subsequent bundles of articles will be published gradually, accompanied by a cumulative expansion of this editorial – a rolling editorial, so to speak.

The contributions of this issue

The article “Making the Whole Greater than the Sum of its Parts. Taxonomy Development as a Site of Negotiation and Compromise in an Interdisciplinary Software Development Project” by Jennifer Edmond, Alejandro Benito-Santos, Michelle Doran, Roberto Therón, Michał Kozak, Cezary Mazurek and Eveline Wandl-Vogt presents the design of a taxonomy of sources of uncertainty in digital humanities datasets with an interdisciplinary and international group. A special focus in this contribution lies on the process of finding a common ground between different communities of practice.

With their article “Visualization of Categorization: How to See the Wood and the Trees”, Ophir Münz-Manor and Itay Marienberg Milikowsky contribute a paper that discusses how figurative language has been annotated in a corpus of Late Antiquity Hebrew Liturgical Poetry – first on paper and then with the annotation and analysis software CATMA – and which new kinds of insight this transition from analog to digital was able to show. The contribution also introduces a way to visualize the annotations in a way that supports category-based hermeneutic speculation about the analyzed texts and phenomena with the visualization tool Vis-À-Vis.

Marlene Ernst, Sebastian Gassner, Markus Gerstmeier and Malte Rehbein contribute an article titled: “From Information Extraction to Reusable Data Models – the Example of the Special Court Munich as Digital Microhistory”. This paper presents three different approaches to categorizing semi-structured information concerning legal history. It discusses the development of a categorisation system for the analysis of approximately 10,000 inventory entries for legal cases from the Special Court Munich (1933–1945).

The paper “From semi-structured text to tangible categories: Analysing and annotating death lists in 18th century newspaper issues” by Claudia Resch, Nina C. Rastinger and Thomas Kirchmair reports the process of developing and applying categories for annotating death lists in 18th century Viennese newspaper *Wien[n]erisches Diarium* and provides examples for the usefulness of the generated metadata in the context of (quantitatively) analyzing topographical and biographical data.

Notes

[1] Both workshops were associated with the project forTEXT; workshop reports can be accessed via the project website: <https://fortext.net/news/2020/workshop-report-non-hierarchical-concept-ontologies-and-markup-schemas> and <https://fortext.net/news/2021/workshop-report-development-and-application-of-category-systems-for-text-research>.

Works Cited



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.