





## Categorising Legal Records – Deductive, Pragmatic, and Computational Strategies

Marlene Ernst <marlene\_dot\_ernst\_at\_uni-passau\_dot\_de>, University of Passau, Department of Digital Humanities   
<https://orcid.org/0000-0003-0097-2267>

Sebastian Gassner <sebastian\_dot\_gassner\_at\_uni-passau\_dot\_de>, University of Passau, Department of Digital Humanities 

Markus Gerstmeier <markus\_dot\_gerstmeier\_at\_uni-passau\_dot\_de>, University of Passau, Department of Digital Humanities  <https://orcid.org/0000-0002-4283-6434>

Malte Rehbein <malte\_dot\_rehbein\_at\_uni-passau\_dot\_de>, University of Passau, Department of Digital Humanities   
<https://orcid.org/0000-0002-3252-0604>

### Abstract

Reprocessing printed source material and facilitating large-scale qualitative as well as quantitative analyses with digital methods poses many challenges. A case study on approximately 10,000 inventory entries for legal cases from the Special Court Munich (1933–1945) highlights those and offers a glimpse into a digitisation workflow that allows for in-depth computer-aided analysis. For this paper, different methods and procedures for developing categorisation systems for legal charges are discussed.

## Remarks on Categorisation – Theoretical Considerations

Many printed historical sources still await systematic reprocessing, or “upcycling” [Donig and Rehbein 2022] in order to facilitate large-scale qualitative and quantitative analyses by digital means. In this instance, we are dealing with approximately 10,000 legal cases from the Special Court Munich throughout the entire period of the National Socialist regime in Germany (1933–1945) which were thoroughly documented in a printed format as “Archive Inventory” in the 1970s [BayStMUK 1975–1977]. While the original court files are heterogeneous in their structure and comprise all kinds of documents, from handwritten notes to typewritten forms, and therefore are not predestined for a systematic and automated approach in extrapolating the content, the documentation done in the 1970s, the “Archive Inventory”, presents us with a semi-structured repository of knowledge that can be analysed by digital methods. First steps towards such were set about recently [Gerstmeier et al. 2022]. For further analyses, we have to pre-process the semi-structured data from the “Archive Inventory” – i.e., categorisation in the form of organising the information contained within the source material. 1

The ultimate goal of the future work is a large-scale study of the sentencing practices of the Special Courts. In this study, for example, correlations between social status, accusation, and sentencing level are to be sought as well as general patterns. Overall, this serves the objective of describing the legal norm (or “average” practice) and at the same time uncovering anomalies (cf. [Schlumbohm and Gribaudo 1998, p. 28]). We are discussing the following six theoretical considerations: 2

1. Categorisation moves between macro- and microhistorical approaches. Such a path to a comprehensive macrohistorical study contrasts with smaller-scale microhistory, for which studies based on few samples of single records already exist. For computer-assisted macrohistorical research, corresponding digital tools are being developed under the metaphor of “macroscopy” [Graham, Milligan, and Weingart 2016].

However, this research approach generally still awaits a theoretical foundation. When Jürgen Schlumbohm [Schlumbohm and Gribaudi 1998, p. 29] discussed whether micro- and macrohistory are to be understood as complementary or incommensurable, he concluded with regard to the aggregation of data following categorisation that “a kind of ‘theory of aggregation’” does not yet exist for historical science. This still seems to be true, and the resulting desideratum needs to be addressed. This is not only the case because of the increase of data-driven research also in “traditional” Humanities disciplines such as History, but especially because a sharp division (in the understanding of a counterpart) between macro and micro studies seems to be as outdated as the rigid adherence to disciplinary boundaries. However, just as there might be no interdisciplinarity without disciplines, the interchange between macro and micro studies should not lead to the abandonment of the cores of the respective research paradigm, but rather form a smooth transition. It is here that categorisation or general modelling forms an important bridge between the study of the particular at the micro level and that of the general at the macro level — from the close-up view to the distant view and back again. Such bridge-building, however, must be purposeful as well as systematic and shall not result in eclectic bricolage and thus arbitrariness.

2. Categorisation is data modelling. All categorisation is modelling and the models we create help “understanding the world by means of patterns and principles” [Bod 2018, p. 78] in the sense of the macrohistory we seek. Hence, general model theory applies to categorisation. Among other things, this means that categorisations are reasoned (and subjective) selection processes. This is of considerable methodological importance for any data-driven work, especially in the Digital Humanities, but is apparently too often neglected, which significantly diminishes the value of the results. Categorisations are conscious typifications or exemplifications and must be demarcated not only from arbitrariness but also from unconscious formation of stereotypes (cf. [Ginzburg and Poni 1991]).
3. Categorisation is complexity reduction. This, too, forms an area of tension between micro- and macrohistory. In the context of methodological approaches of the Digital Humanities, reference could be made here to procedures of close and distant reading. In close reading, the human researcher processes significantly more information from each source but fewer sources, whereas, in distant reading, significantly more sources are condensed into less information. In our case, for example, microhistory corresponds to the close reading of complete, more substantial, but only few sample records, while the macrohistorical approach corresponds to the distant reading of aggregated data. In such a reduction of complexity, it has to be weighed up how much of the source content has to be taken over 1:1 and how much may be typified. The higher the degree of abstraction, the greater the danger of losing relevant content without achieving any more selectivity. This is a problem we have encountered especially in the pragmatic approach below. In the case of the Special Courts, reducing potentially substantial sources (“*Akte*”/file) to a few data (“*Aktenregest*”/record), which are then additionally aggregated into categories, could be considered typical for macro-historical questions based on secondary data from “upcycling” processes.
4. Categorisation of historical data means dealing with time. The period of validity of categories is often limited, which becomes especially evident over the transition from pre-modern to modern times (“*Sattelzeit*”) but should be seen as a fundamental challenge of historical macro studies in general. Here we usually have two levels to consider: first, the present time from which we write history, and second, the past about which we write, whose sources we use, and whose terminology we rely on. In the case discussed here, there is even a third temporal level, namely the indexing of the files carried out in the 1970s and the vocabulary used there to produce the “Archive Inventories”. This results in the challenge to do justice – as far as possible – to the conceptuality and the vocabulary of the source while at the same time creating a valid categorisation over different temporal levels and evaluating it critically (cf. [Seiffert 2006, p. 205]; [Wehler 1972, p. 20]).
5. Categorisation is hermeneutical and ontological. It has already been described in the literature that there is a tense but also distinctly appealing relationship between source vocabulary and modern typology. In addition, in our case it is especially important not to fall for the Nazi ideology and the language they used (cf. [Klemperer 1947/2020]), which often stands in contrast to actually valid law (of the Weimar Republic), to a modern understanding of law, and to human rights, which are considered universal today. The question of

universals alluded to here is also of fundamental importance for categorisation. Seiffert [Seiffert 2006, p. 159] distinguishes between “source terms”, “modern terms”, and “supra-temporal terms”. Certain terms, he claims are “timeless”, such as the category “human being,” and thus “so general and formal that we can unhesitatingly apply them to any historical epoch without fear of doing violence to the historical fact”. It is essential to keep this ontological dimension in mind when categorising. In the three approaches we have followed, especially the inductive-computational (automated) one attempts a clustering via the source terms (of the secondary source). It takes the finding of the source terms without further interpretation as its basis of categorisation, while the first two approaches interpretively process a priori or inductive theoretical presuppositions. But here, too, the question remains: is this interpretation present-tense, is it situated within the hermeneutically understanding of the past under investigation, or can it be supra-temporal?

6. Categorisation has a normative function. Directly linked to these considerations is the question of whether a categorisation is descriptive or normative: does it describe (ontologically: “A is an X”) or does it norm (“shall A be an X”, “shall A be understood as an X”)? A source approach is arguably more descriptive, while the use of modern terms tends toward a normative view. Thus, categorisation, like any kind of norm-setting, runs the risk of having a domination or even ruling function inherent in it (cf. [Richter 2021, p. 53] with reference to [Durkheim and Mauss 1901]). This is particularly relevant in data-based research: Whoever prescribes the data categories sets norms. Therefore, given categories must always be critically questioned, which is especially important when “upcycling” older data. This also applies to primary data (take Census data, for example: here the exercise of domination is particularly conspicuous, which admittedly was also practised under Nazi rule, but not only there, see e.g., race laws). Such categorisations can then have an impact on the formation of society/ies.

The following pages present a case study on three different approaches towards the categorisation of legal sources from the Nazi era taking three different points of view into account: a) a deductive one, which operates on the basis of legal-philosophical and legal-historical theory; b) a pragmatic-explorative one, operating with secondary source texts, but using contemporary understandings for category formation; and c) an inductive-computational approach. They highlight the varying spectrum and different views one can take on the past while simultaneously showcasing the interplay between the theoretical considerations described above.

3

## Introduction to the Source Material

Although not an entirely new phenomenon in NS Germany, the Special Courts were distinctive in the sense that they combine regular criminal law and charges of the form of non-conformity according to Nazi ideology under one jurisdiction. Not least because of this, the contents of the cases are particularly diverse and hard to grasp in modern legal terms. The corresponding case files are just as complex and heterogenous in their content and structure – from typewritten forms to handwritten notes. The compilatory work done in the 1970s helps to get a better grasp on the content of the original files by putting the information in a model of a more concise and structured nature.

4

The inventory, which was compiled by a team from the Bavarian State Archives between 1975 and 1977, is in itself remarkable in several respects. It was printed in a limited quantity of copies, self-published in off-set technology, bound off-site and not distributed on the book market, but given away to libraries and archives. The resulting volumes are thus rather to be understood as “grey literature”, as auxiliary means for the internal use of authorities and for scientific purposes. They are not only a finding aid but a possible answer to the challenge of indexing mass sources and a source in themselves, reflecting editorial decision-making processes of that time. In the example of entry number 165, we get a glimpse of the overall structure of an inventory record with the main content description by the authors in English added in blue:

5

occupation: driver      name      date of birth  
 Prozeß gegen den Kraftwagenführer Johann SOYER (geb. 3. Dez. 1905),  
 personal/political background  
 (optional):      accusation:  
 SA and Nazi Party member      residence: Munich      spreading the rumour  
SA- und NSDAP-Mitglied aus München, wegen Verbreitung des Gerüchtes  
about the champagne party in the Brown House  
über das Sektgelage im Braunen Haus.  
 result of the proceedings: 2 months prison sentence  
 Urteil: 2 Monate Gefängnis  
 duration of the proceedings  
 9. Jul. 1933 - 12, Spt. 1934  
 registration number  
 (S Pr 213/33)

Figure 1. Example of inventory record number 165 with added content description in blue.

Instead of highlighting individual fates and putting emphasis on particular information of the respective procedure reflected in the file and its transmission process, the editors decided to define a standardised set of general characteristics to be collected for each case. That this standardisation changed over the long-stretched process of compilation is only natural, e.g., for cases from 1939 onwards, the underlying relevant legal paragraphs have been added. Thus, additionally complicating the computation process of today. As researchers of digital history, this circumstance confronts us with several challenges at once, especially in the context of categorisation.

Nonetheless, analysis of a more comprehensive nature is made possible through the digital transformation. Because the corresponding primary sources, i.e., case files, have been analysed only partly, the inventory still retains great relevance for research today. The main aim lies in digitally upcycling the inventory in order to make the records available and pre-process the data for quantitative analyses, e. g., correspondence analysis for which well-designed categorisation is essential. After digitisation and pre-processing, information extraction from the machine-readable text is possible through its formalised structure. The compilatory work done in the 1970s has been meticulous and offers access to the content on a systematic level. Nonetheless, close reading raises questions about how the editors structured their approach. There also seems to be some inconsistency in how the information was processed (e.g., the wording of the charges laid against the defendants). Those inconsistencies pose a challenge for the digital upcycling in the sense of automated processes taking, for example, the extraction of accusation categories: For further large-scale analysis, cases on derogatory or critical remarks – by far the largest group of charges – belong in the same category, whether they may refer to the spreading of rumours (on Hitler or otherwise) or rants about the regime in general. Those cases are to be differentiated from crimes still liable to prosecution today, like rape or theft.

## Workflow – From Paper to Dataset

Automatic recognition of typewritten text (OCR) is a comparatively established field that already has a number of standardised procedures. Permissively licensed OCR engines such as Tesseract [Smith 2007] or Calamari [Wick, Reul, and Puppe 2018] are not inferior in quality to proprietary solutions [Reul et al. 2018]. The workflows associated with these solutions can be better shared and adapted for similar usage contexts. As part of the project, we therefore implemented and incrementally improved numerous pipelines.

Initially, the books have been scanned using an overhead book scanner. Unsurprisingly, the resulting digitised pages showed warping and skew. Several attempts of performing optical character recognition after automatic de-warping and de-skewing did not lead to satisfying results. The overall error in the OCR results introduced by distorted pages, combined with deteriorated and contaminated typeface could not be reduced to an acceptable level, after the fact, by software alone. Page warping is due to the pages being bound into books. By unbinding the books and digitising single pages using a document feeder scanner, page warping and page skew could be eliminated almost completely. Deterioration and contamination of the typeface were still present, but the overall error rate was significantly reduced.

9

In the next step, we employed Tesseract for OCR on all pages, specifying a German dictionary and the calligraphy hand “Fraktur” and outputting one file per digitised page, in hOCR format [Baierer 2020]. Compared to plain text, hOCR has the added advantage of preserving layout information. Each page contained several inventory records, with entries sometimes spanning two pages, when starting close to the bottom of a page. We continued processing the text programmatically by undoing the physical boundaries at page-level and, instead, splitting the text at logical boundaries, namely at the inventory record level. Out of 9,955 inventory records, 8,531 or 86 % could be extracted successfully [2]. With separate inventory records in plain text format, our algorithm continued to parse and translate each entry into JSON format [3]. Each JSON record captures meta information, such as page number and name of the hOCR file containing the record, and the content of the inventory record itself in separate fields, as detailed in Figure 1. It should be noted that we also retain the underlying plain text within each JSON record, to maintain the opportunity to make (manual) corrections or to retrace decisions later.

10

In a final step, we selected relevant information from a subset of JSON fields and created a CSV file, serving as the baseline for the different categorisation approaches described throughout the rest of this article. Each row in the CSV file contained a unique identifier, the inventory record, the “law” or legal basis / “*Gesetzesgrundlage*”, respectively, and the duration of the case, as well as normalised versions of the inventory record and duration. The normalised inventory records retained only the accusation / “*Anklagegrund*”, stripping off the result of the proceedings and lemmatizing all words. The normalised duration contains the starting date only, standardised as YYYY-MM-DD. After this pre-processing, the information is principally available for computer-aided analysis but additional strategies for systematically approaching the data had to be developed.

11

Different approaches may be taken in processing, categorising and engineering an in-depth analysis of the 10,000 cases. After the initial digitisation and information extraction – processes which were formalised with regard to reusability and a proof-of-concept workflow for source digitisation in mind –, challenges are posed by different categorisation models that may be applied to the corpus. The different methods, from an intuitive approach based on keywords to automated processes utilising natural language processing and clustering the results, exemplify how varying methods can influence our view on the past. In this paper, these different perspectives and their interconnections will be discussed in order to open up a new scholarly approach to the inventories. They are especially important for subsequent statistical analyses.

12

## Categorisation Methodologies

Even though positive law was seen critically or even derogatory within the ideologically affected legal system of the “Third Reich”, NS judiciary somehow cleverly understood to use the benefits of a traditionally grown and reliably functioning judicial machinery like in modern Germany for its own ideological means of arbitrariness and terror throughout all changes of 20th century political systems [Lahusen 2022]. The other way around, the extent to which the German judiciary more or less voluntarily made itself a “stooge” ([Weber and Piazzolo 1998, pp. 13–14]; [Gruchmann 2001]) of the Nazi regime is exemplarily proofed by the Special Courts, while their specific legal bases, which were enacted from 1933 onwards without any parliamentary participation, exemplify the dissolution of the legislature in the “Third Reich”.

13

Simultaneously, the ancient Greek etymology of the term “category” comprises primarily nothing else than a legal significance – “*κατηγορεῖν*” means “to accuse”, “to charge” ([Gemoll 2006, p. 453]; [TLL 1906–1912, v. III, col. 602]. As we focus on a court, a categorisation along its authoritative working base regards a deepened consideration of

14

applicable laws, legal regulations, ordinances, decrees, or even – depending on the granularity depth of data modelling – the specific legal paragraphs that were valid at the time. And indeed, each decision of a court in the modern sense refers to a concrete legal basis: Even in the “*Unrechtsstaat*” of NS dictatorship judges could not simply render judgements at their own discretion, according to the legal principle “*nulla poena sine lege*” [Feuerbach 1801, p. 20]. This formal aspect of Special Courts was dedicatedly criticised by ideologically convinced Nazis in German Judiciary and Faculties of Law ([Garbe 1999, pp. 141–142]; [Fröhlich 1983, p. 210]). So, purely theoretically, such a deductive approach might be the most reasonable.

## a) Deductive-(Legal-)Philosophical and (Legal-)Historical Approach

Our first approach of categorisation can be classified as *deductive*. There is a certain degree of legal-philosophical background behind the historical phenomenon investigated, but more prominently we are offered references to specific legal sources – at least for part of the case files. 15

Deduction is considered a most distinguished philosophical method of categorisation [Kant 1787/1904, pp. 104–130, §§14–27]. The most paradigmatic modern contribution to the conceptualisation of deduction as a philosophical method relates to law [Fichte 1966/70]. Concerning the NS Special Courts, a categorisation alongside the legal bases is particularly interesting as their qualities are remarkably diverse. First of all, we chose a legal theoretical distinction between the positive law [Kelsen 2017] and laws and legal regulations, respectively, in the sense of Carl Schmitt’s [Schmitt 1922/2021] and other NS legal thinkers’ theory of law and society. It marks the peculiarity of NS Special Courts within the contemporary judicial system of the Reich that in the same way of their institutionalisation beyond the regular courts, the NS regime began with its legislation of special penal regulations besides the regular criminal laws that had been systematically codified in the German Nation State since 1871 [RStGB 1871, 1935]. The proceedings before the Special Courts concerned first of all exclusively such “crimes”, or – to be more semantically neutral – actions that only had been criminalised by a certain kind of NS special penal regulations. A closer look on the complex process of the Nazi Seizure of Power shows that the NS regime relaunched special penal legislation even before relaunching the Weimar-invented Special Courts [Verordnung 1933], at first by means of the “Reichstag Fire Decree” [Reichstag Fire Decree 1933]. The scientific estimation of this ordinance as “the fundamental exceptional law on which the National Socialist dictatorship was primarily based until its collapse” [Bracher 1960, p. 82] shows the principal significance of this kind of repression for the establishment of NS dictatorship. 16

*D-1:* The most important and in over twelve years of NS special penal judiciary most often applied special legal regulation [Hüttenberger 1981] became the so-called “*Heimtücke-gesetz*” [Heimtücke-gesetz 1934]. Thereby, the term “*Heimtücke*”, i.e., “insidiousness”, must be understood as a typical NS dysphemism. This law criminalised every kind of verbal or written critical comment on the NS regime or on certain Nazi politicians – whereas in regular German criminal law, “insidiousness” is actually an essential characteristic of murder [RStGB 1871, 1935, §211,2]. We define the early NS special penal law regulations as our *category D-1*. 17

*D-2:* As another facet of their peculiar role within the NS system of persecution and repression, Special Courts were increasingly used for “regular crimes”. The longer the NS regime lasted, the more frequent such cases were brought before it. This phenomenon is well known by existing research literature (e.g., [Dörner 1998, pp. 36–39]; [Wogersien 2007]). Apart from a few exceptions (e.g., [Graczyk 2021, pp. 250–330]), concrete numbers or data on this topic have not been presented yet. The Mannheim Special Court shows, for example: In 1933 18.6 % of “regular” first instance judgements (like theft or murder) were tried before it, but in 1943 this percentage would increase up to 43.9 % [Oehler 1997, pp. 36–37]. Precise statistical analyses on the question of how often in 1933 still valid laws [RStGB 1871, 1935] were applied represent a desideratum. By the example of Special Court Munich, this can be approached through our computer-based investigation of its archive inventory from the 1970s (cf. section below). Besides the codified law, Munich Special Court even applied rather marginal regulations and decrees of regular penal law for its judgements, up to the late Weimar “Animal Protection Act” [Animal Protection Act 1930] or exclusive Bavarian state laws – what surprises in the face of anti-federal Nazi system –, like the “Slaughter Tax Act” [Slaughter Tax Act 1930]. 18

*D-3:* Research so far has also thematised that the beginning of World War II meant a significant incision in the history of 19

NS special judiciary. Already since 1938, a veritable boost of new specific NS special penal law regulations began to unfold. This “second wave” of NS special criminal law-making is correlating with the sheer increase in the number of cases brought to trial before, e.g., the Munich Special Court since 1939: Between 1933 and 1938, 3,123 proceedings were carried out, but in the years 1939–1945 they reached 6,823. This tremendously increasing repression no longer exclusively referred to the regime’s “own” German “*Volksgemeinschaft*” but to foreigners (mostly forced labourers and prisoners of war), too, since the annexations of 1938 and 1939 and particularly since Germany’s war of aggression. This justifies the formation of another deductive-legal-historical category, especially since these new products of NS special penal law regulation [4] – in contrast to the more generally applicable “*Heimtücke-gesetz*” – are concerned with more or less specific war-related deviance or resistance, or with the racial or ethnic affiliation of prosecuted people itself, respectively:

- The most universally formulated regulation of D-3 is the “*Volksschädlingsverordnung/VVO*” [Volksschädlingsverordnung 1939]. With its typical dysphemic Nazi-language title [Klemperer 1947/2020, pp. 291, 403] and enacted four days after the beginning of World War II, Munich Special Court would apply it 1,074 times until 1945, often under the imposition of the death penalty, which VVO §1 provides, e.g., in the case of looting. The “*Wehrkraftverordnung*” [Wehrkraftverordnung 1939] criminalised actions that, in the eyes of the NS regime, undermined the defence force. By 1945, Munich Special Court had applied it 779 times.
- Beyond these, the NS regime enacted at the very beginning of World War II special penal regulations concerning specific kinds of resistance in the context of everyday life during the war, too, such as listening to hostile radio programs [5] (cf. [Hensle 2005]; [Christians 2020, pp. 265–276]) as well as a whole range of ordinances dealing with individual misbehaviour to the harm of war economy ([Christians 2020, pp. 217–227]; [Hackl 2022]), e.g., the “*Ordinance on War Economy*” [Ordinance on War Economy 1939], applied 599 times by Munich Special Court.
- NS special penal law regulations and judgment routine of the Special Courts were over time even involved in the ideological and racist war of determination and the Shoah, which is manifested especially by the “*Penal Ordinance against Polish and Jews*” [Penal Ordinance against Polish and Jews 1941].

D-4: Finally, in the context of “total war”, the judicial practice of NS Special Courts reached out even to the legal basis of another non-regular jurisdiction instance, namely to the proprietary penal law of the German “*Wehrmacht*”. The Military Penal Code [MStGB 1943], originally launched by early *Kaiserreich*, would be the legal basis for three judgments of the Munich Special Court in 1944/45. 20

For an understanding of the historical phenomenon of NS Special Courts and their – since 1939 once again changed – role within the Nazi repression system, a deductive categorisation along the concrete legal bases of proceedings may provide a substantial epistemological profit. It might even go to the core of the complex relationship between National Socialism and law as well as justice, respectively, in a diachronic perspective. As far as the legal bases of the Munich Special Court during World War II are concerned, we are dealing with a simultaneous set of legal sources that arose at different times and, according to legal theory, are to be evaluated differently. This multimodality of legal bases, some of which are applied simultaneously in proceedings, is to be modelled in categories D-5, D-6, and D-7 (concrete numbers cf. below, “*Comparison*” and Figure 5). Especially since 1939, Munich Special Court often would apply D-2 legal bases – above all paragraphs of the RStGB [RStGB 1871, 1935] – in addition to the D-3 NS special penal law regulations whenever possible. Also, the “*Heimtücke-gesetz*” (1934) of D-1 would still be applied extensively in addition to the new special penal law regulations from 1938 onwards; even the “*Reichstag Fire Decree*” (1933), still kept being applied by the Munich Special Court, even in times of “total war” (17 times, remarkably in most cases against Jehovah’s Witnesses). 21

Ideally, deductive categorisation seems to be the most rational method of understanding the actual nature of this historical phenomenon. In our present research, however, it indeed proves to be only partially purposeful. Formally, the façade of a rationally acting judiciary was maintained even under the conditions of “total war” [Lahusen 2022, pp. 141–185], illustrated with the example of Aachen Special Court), although the Munich NS Special Court was undoubtedly part of a state of injustice. The problem lies rather in the fact that in our research we do not work primarily on the basis 22

of the court records themselves but with the help of secondary records prepared in the 1970s (cf. above, introduction to the source material). As is the case with many digital humanities projects today, the editors of the “pre-digital” database of the BayStMUK-“Archive Inventory” changed their metadata scheme in the course of data entry. Thus, it happens that the laws/ordinances/legal paragraphs that had been applied by the NS Special Court were only recorded in the inventory records from the beginning of Part 3 (1939) onwards during the scientific retrospective of the 1970s [BayStMUK 1975–1977, 1976, p. 619; 1977, p. 2349]. In the records on the trials and discontinued proceedings of the years 1933–1938, no legal basis is given in the records of the “Archive Inventory” at all and for the phase from 1938 to 1945 are also some instances where such explicit information is missing (specifically in 473 cases). Nonetheless, even for the proceedings where the record entries offer no paragraph in the “Archive Inventory”, the offence recorded in each case can always be used to draw conclusions about an unambiguous criminal offence and the applied written law (cf. below, comparison). For the research design of our project, this means that a categorisation along the legal basis of the special court proceedings actually only makes sense for the period from the beginning of 1939 onwards, even though 6,830 of the entire 9,954 cases captured by the “Archive Inventory” took place after the beginning of 1939.

For our research and the formation of meaningful and, above all, useful categories for the quantitative historical analysis of the (secondary) source corpus, this lack of data and missing information means that we have to find a more pragmatic methodology in order to focus on the actual object of investigation – i.e., the Munich NS Special Court and its repressive practice or the “everyday” resistance and deviance of the wider population in the “Third Reich” and the “wartime society” ([Wildt 2007]; [Wildt 2019, pp. 302–325]) that manifested itself before this court.

23

## **b) Pragmatic-Explorative Approach**

Through a random sample of cases for close reading and the intensive study of the source material in the context of the digitisation and data modelling process, we could gain an overview of the content. This insight allowed us to think of several topics relevant to categorising the charges by applying an intuitive system. In some cases, the demarcation was relatively easy: A huge part of the charges concern negative remarks against the regime (e.g., the claim that the Reichstag was set on fire by National Socialists) and leading figures thereof (e.g., the allegation that Hitler and Göring are gay). Another very clear category deals with social interactions with prisoners of war. Non-war-related inventory records contain crimes from bodily harm, e.g., in the form of murder, to theft. Evaluating the charge content showed that the wording used to describe the individual accusations is in many cases standardised. This offered the possibility of sorting the whole corpus with a few selected keywords per category – so our hypothesis.

24

A first trial run of the approach – by defining eight categories, by analysing a sample of 90 randomly chosen cases and looking for similarities in their content for classification (1-bodily harm, 2-fraud, 3-economic crimes, like theft, 4-war-related economic crimes, like illicit butchering, 5-interactions with prisoners of war, 6-sophisticated regime critique, written and oral resistance, 7-conspiracy, like pamphlet distribution, 8-negative remarks) – revealed a disproportional surplus with over 40 % of cases concerning negative remarks. To cater for deeper statistical analysis in the future, differentiating within this main part of accusations became relevant in order to gain better insights into the relationship between accusations, results, and socio-political backgrounds of the accused. So, through a review and quality control loop, we decided to refine categories, what they each comprise as well as which keywords to use. Thereby, comparing the success and failure of predicted results from single keywords places induction in a pragmatic context (cf. [Holland et al. 1986]).

25

The query or category allocation was carried out on the basis of the normalised charge text. The number of keywords necessary per category varies. Through a review loop for each, the most efficient and concise word variations were sought and found. Working with a spreadsheet, the course of action was to test keyword variations on the column with the normalised charge text before incorporating the best version into a nested formula, i.e., the most efficient one in terms of the number of cases that could be allocated without compromising the results for other categories. In some cases, only parts of words and phrases were used in order to incorporate as many different word and verb forms as possible without compromising the results by accidentally returning cases dealing with wholly different content. This process led to the following categories with the different topics included in each.

26



*E-1* bodily harm and morality: The primary aim was to separate cases of murder, manslaughter and other bodily injury offences from those that were purely regime-related. A review loop revealed that many charges deal with people accused of talking about how Hitler and other NS functionaries should be murdered or otherwise harmed, but where no actual deed was done. Additionally, the review process revealed that there were also many instances of morality (indecent behaviour) and rape brought before the Special Court. Though many different accusations are part of this category and almost 20 keywords were necessary for allocating them into this category (the highest number from all out of the categorisation process), overall, only 264 instances are dealing with cases that are still (mostly) relevant under criminal law today.

*E-2* fraud: 308 cases are related to fraudulent behaviour. Only five keywords sufficed in order to collect all relevant instances. “*Betrug*”/fraud was commonly used in the inventory in this case but also forgery (e.g., of documents) gets mentioned.

*E-3* theft and bribery: This category comprises not only economic crimes still punishable by law today, like theft of goods, burglary, or tax evasion but also instances where begging or embezzlement are concerned. In the end, 909 cases are sorted into this category.

*E-4* war-related economic crimes: Cases in this category are differentiated from the former by their relation to war. They mostly concern food supply whereby black-market dealings, illegal slaughtering of livestock, and illicit dealings with ration coupons are meant. Additionally, foreign exchange crimes are also put in the same category which, in the end, comprises 644 cases.

*E-5* prisoners of war-related crimes: This is another instance, similar to the second category, where sorting was relatively easy and a small number of keywords sufficed to cover 661 cases. The wording, for the most part, stays the same throughout all inventory records for the cases where – mostly (young) women – helped or socially interacted with prisoners of war. Additionally, escape aid was also put in this category.

*E-6* regime criticism including religious and political persecution: Differentiating between categories 6 and 7 involved the toughest decisions in the process. In the end, mainly due to the extensive review and quality control process, it seemed most prudent to aggregate political and religious persecution (like members of the Communist Party and Jehovah’s Witnesses) with instances where politically or religiously motivated active resistance in the form of written statements or preaching against the regime during Sunday service are concerned. Initially, cases involving the unauthorized ownership and/or wearing of NS-uniforms or badges were also put here. But the main motivation behind this pretence of being part of or supporting the regime was self-preservation, thus not being on the same resistance level as the other entries within the category. Those keywords were therefore relegated to category E-7 and also resulted in decimating the number of relevant cases by approximately 200 till only 238 remained in category E-6.

*E-7* conspiracy and protest: Through the additional cases moved from the former one, 528 instances could be found where active resistance in the form of protest or consummation of illegal media but also active actions for self-preservation are concerned. This also comprises the distribution of leaflets as well as listening to foreign radio channels.

*E-8* negative remarks and hearsay: Although the review and quality control process helped in distinguishing more properly between the different categories, the by far most extensive category remains the one related to negative remarks. 3,005 instances can still be ascribed to people commenting on different aspects of the regime in different settings – from people repeating hearsay about deaths in concentration camps to others insulting the government during a session at the local inn.

*E-9* person-related remarks: As the first trial run of this approach showed, disproportionately many cases deal with accusations based on negative remarks and/or hearsay. In order to differentiate those a little bit more, cases where Hitler/the “*Führer*” or other leading figures of the regime, like Goebbels or Himmler, are part of the claim are separated from the rest, resulting in 1,523 cases.

Due to varying phrasing of the editors from the 1970s for the charges within the inventory records (e.g., for negative remarks we have “*schimpfen*”/ranting, “*negativ, abfällig äußern*”/speaking derogatorily, “*kritisch äußern*”/critical remarks etc.), partly due to the long process of collecting data and creating the inventory and partly due to the fact, that different persons worked on them as well as still remaining OCR-errors, it helped to keep the single category keywords as short

as possible. Thus, we encapsulated varying spellings and verb forms as well as common OCR-errors like “a” instead of “ä” within one query. As the search for the keywords is operated via the normalised and shortened text, which is sometimes very individual in its wording concerning the charges, additional approximately 100 cases, which otherwise would result in being relegated to the group of uncategorised records, can be allocated to the group of “negative remarks and hearsay” by querying for “*gesagt haben soll*” (reported to have said) in the full text at the end. In the end, 451 records remain uncategorised (*E-0*) due to poor quality of the OCR-text or because an accusation related to certain allegations is mentioned only once, e.g., there is one accusation of poisoning wells, a trope used against Jews since the Middle Ages.

The resulting formula of nested if-then queries [6] is quite intricate but also streamlined concerning efficiency. Nonetheless, one has to keep in mind that the sequence within the query influences the result. If we, for example, retrieve cases by filtering for charges of the category “bodily harm and morality”, like fornication/“*Unzucht*” or murder, first, we would also put cases in the same category where hearsay in relation to such crimes is concerned. The assigning of the cases to single crime categories by particularly chosen keywords is subject to individual biases as some – like the unauthorised wearing of badges and uniforms (ultimately put into *E-7*) – could be placed in different categories depending on where we as editors choose to put them. Therefore, we also developed a more “objective” inductive-computational approach.

28

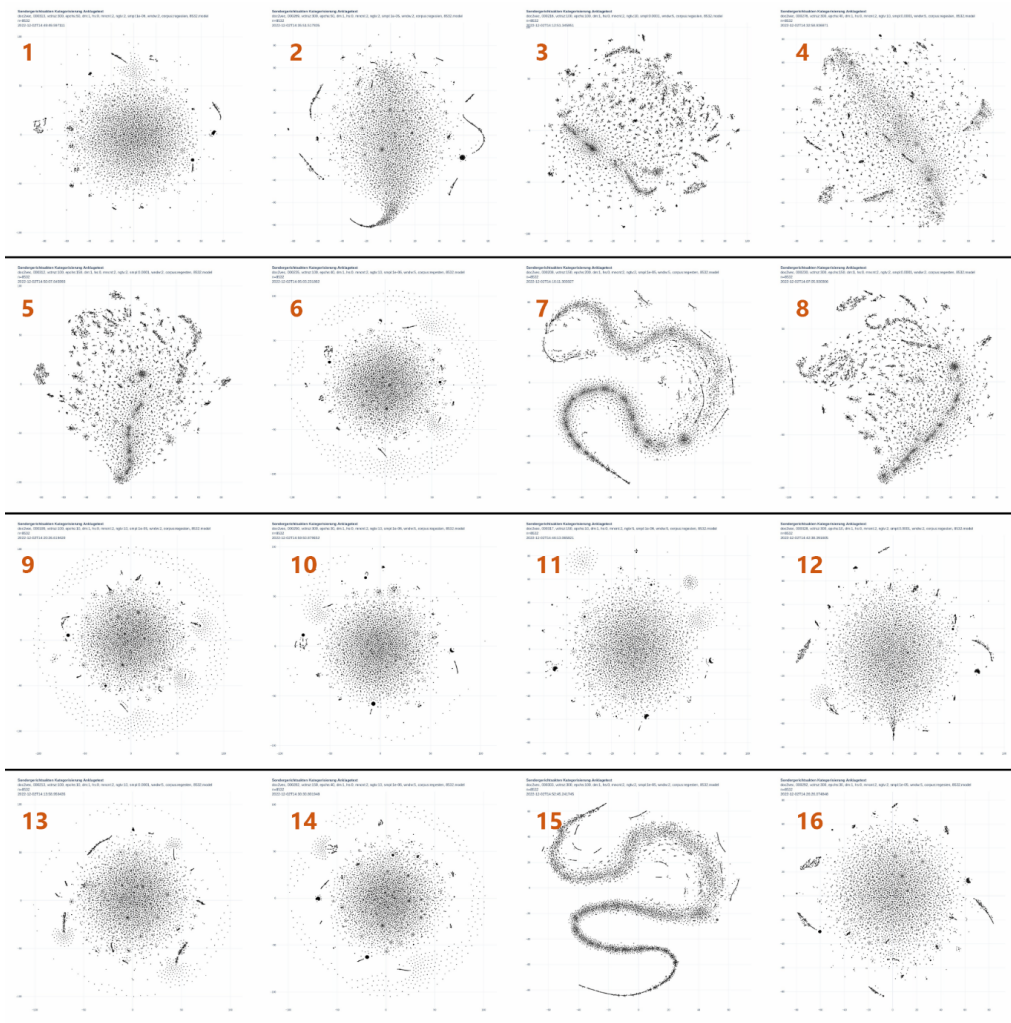
### c) Inductive-Computational Approach

The inductive-computational approach aims at producing a categorisation with little manual intervention by applying algorithms from natural language processing (NLP), machine learning, and clustering. Most algorithms used must be parameterised which requires expert knowledge, structured exploration of the parameter space and experimentation.

29

Starting from the baseline CSV file (cf. “Workflow” above), we first trained a doc2vec model [Le and Mikolov 2014] using the gensim library for Python 3 [Rehurek and Petr 2010]. The algorithm defines an ample set of parameters, some of which have a strong influence on the quality of the document embeddings produced. Our goal was to cluster similar inventory records close together and far apart from non-similar ones. Finding the right set of parameters requires trial and error: define a parameter configuration, train a doc2vec model, obtain embeddings for all documents, and evaluate the results visually. At this point, we would like to note that we added a manual intervention: doc2vec does not solve a classification task (there is no ground truth) but, instead, generates document embeddings. Thus, no metric exists to assess the quality formally and automatically in terms of classification accuracy of the embeddings produced. Therefore, we selected a doc2vec model manually: we first computed embeddings for all documents and plotted them in 2D space separately for each model; we then used visual inspection to select a model that would generate a non-homogeneous distribution of documents with clear clusters. Figure 2 exemplifies how we explored the parameter space to select a model: each subplot represents a model trained with a distinct set of parameters and shows the document embeddings generated by the model projected into 2D space.

30



**Figure 2.** Exploring the parameter space to select a model: each subplot represents one model trained with a distinct set of parameters and shows embeddings for all documents after projection into 2D space. For example, the models 9–12 show a homogeneous distribution of documents, without pronounced clusters; other models, such as model 3, are better for our purpose because they result in more structured alignments of document embeddings.

Even a small set of parameters result in a huge space to explore: for example, five parameters with three possible values each would require  $3^5 = 243$  models to be trained and assessed. To limit the number of parameter configurations and, thus, the number of models to be trained, we relied on our expertise and followed the advice given by Lau [Lau and Baldwin 2016] on recommendations for optimal doc2vec hyper-parameter settings. We then systematically explored the parameter space, by first defining for each parameter a range of allowed values and then iterating each possible combination of all parameters. For example, when allowing  $a = \{1, 2, 3\}$  and  $b = \{4, 5, 6\}$ , we would obtain nine possible combinations of  $a$  and  $b$ , namely  $\{1, 4\}, \{1, 5\}, \dots \{3, 5\}, \{3, 6\}$ . After training only a small number of doc2vec models and inspecting the results visually, we were able to identify some parameter values which would always produce insufficient results, independently of all other parameters. For example, disabling negative sampling or switching on hierarchical softmax would always produce a homogeneous distribution of documents, without distinct clusters, which is undesirable for our purpose. We excluded these parameters from the exploration by setting a fixed value and shrinking the parameter space to explore by orders of magnitude.

31

Besides exploring doc2vec parameters configurations, the selection of a training corpus has a significant impact on the quality of the models produced. Initially, we trained our model on our own inventory records and used the same set of documents to obtain document embeddings. Our corpus consisted of 8,532 valid inventory records with an average length of 5.3 words. We were able to improve our models to show more pronounced clusters by using a different corpus during training, containing 1,454 legal documents from between 1933 and 1945 with an average length of 444 words [7].

32

After training and manually selecting a doc2vec model through visual inspection [8], we were ready for the final step: clustering. First, we inferred a high-dimensional embedding for each inventory record and then projected each 300-dimensional vector into 2D space by applying t-SNE [van der Maaten and Hinton 2008]. We then transformed our data into quantiles [9] before clustering by applying DBSCAN [Ester et al. 1996]. We used pandas [McKinney 2010], numpy [Harris 2020], and sklearn [Pedregosa 2011] to implement all steps. The image below shows a clustering of document embeddings representing our inventory records.

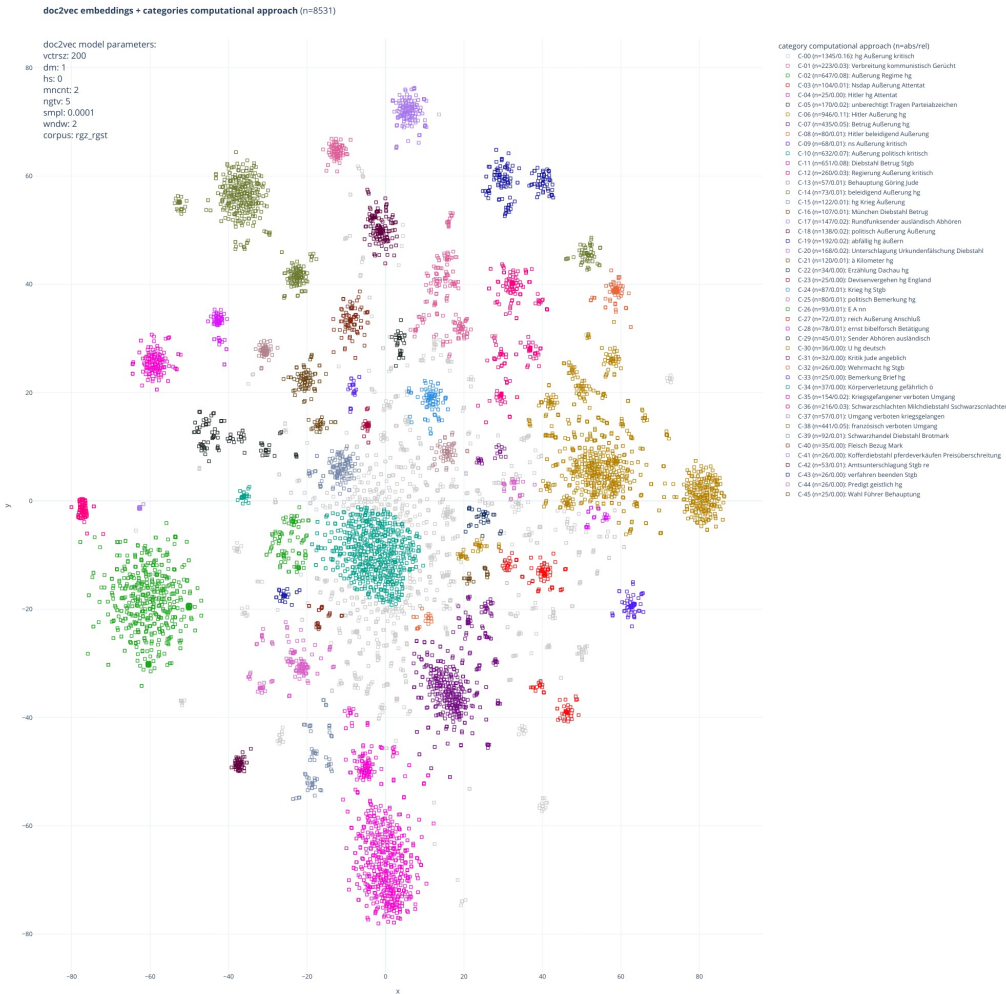


Figure 3. Document embeddings, clusters, and categories from the inductive-computational approach.

Each dot represents a doc2vec embedding which corresponds to a single inventory record. Similar documents are located close together and form clusters. Each cluster corresponds to one category. Outliers are not part of any cluster. Note that the image shows results from a particular choice of parameters used for training the doc2vec model and for clustering document embeddings. We manually chose a set of parameters. Different parameters would produce a significantly different output.

The number of clusters/categories in this approach was much higher than in the other two approaches. This was because documents were grouped solely based on contextual information found in the texts, whereas the other approaches grouped legally related documents together, even if they were appearing in completely unrelated contexts. For example, category E-5 (prisoner of war-related, pragmatic-explorative approach) collects all accusations concerning contact with prisoners of war, whereas the three categories C-35, C-37, C-38 are produced in the inductive-computational approach, seemingly distinguishing between the nationality of the prisoners.

The following table (Figure 4) shows statistics for all 46 clusters sorted by the number of documents within each cluster.

The third column shows the relative number of documents contained within each cluster. The right-most column shows the top three most frequent words occurring within the cluster. Almost 16 % of all documents belong to C-00 which collects outliers and corresponds to the “uncategorised” category in the other two approaches. The clusters are unevenly sized. The eight largest clusters (17 % of all clusters, not including C-00) contain 50 % of all categorised documents. Out of these eight clusters, five contain the word “Äußerung” (remark) within the top three cluster words which correspond to 35 % of all documents. This clearly shows that a large proportion of all accusations concerned remarks about the regime, leaders, the war etc.

	#	Cluster size		Top-3 words within cluster	
		rel	abs		
1	C-00	1345	15.77%	hg Äußerung kritisch	no category/outliers
2	C-06	946	11.09%	Hitler <b>Außerung</b> hg	50 % of all documents
3	C-11	651	7.63%	Diebstahl Betrug Stgb	
4	C-02	647	7.58%	<b>Außerung</b> Regime hg	
5	C-10	632	7.41%	<b>Außerung</b> politisch kritisch	
6	C-38	441	5.17%	französisch verboten Umgang	
7	C-07	435	5.10%	Betrug <b>Außerung</b> hg	
8	C-12	260	3.05%	Regierung <b>Außerung</b> kritisch	
9	C-01	223	2.61%	Verbreitung kommunistisch Gericht	
10	C-36	216	2.53%	Schwarzschlachten Milchdiebstahl Sschwarzschlachten	
11	C-19	192	2.25%	abfällig hg äußern	
12	C-05	170	1.99%	unberechtigt Tragen Parteiabzeichen	
13	C-20	168	1.97%	Unterschlagung Urkundenfälschung Diebstahl	
14	C-35	154	1.81%	Kriegsgefangener verboten Umgang	
15	C-17	147	1.72%	Rundfunksender ausländisch Abhören	
16	C-18	138	1.62%	politisch Äußerung Äußerung	
17	C-15	122	1.43%	hg Krieg Äußerung	
18	C-21	120	1.41%	ä Kilometer hg	
19	C-16	107	1.25%	München Diebstahl Betrug	
20	C-03	104	1.22%	Nsdap Äußerung Attentat	
21	C-26	93	1.09%	E A nn	
22	C-39	92	1.08%	Schwarzhandel Diebstahl Brotmark	
23	C-24	87	1.02%	Krieg hg Stgb	
24	C-08	80	0.94%	Hitler beleidigend Äußerung	
25	C-25	80	0.94%	politisch Bemerkung hg	
26	C-28	78	0.91%	ernst bibelforsch Betätigung	
27	C-14	73	0.86%	beleidigend Äußerung hg	
28	C-27	72	0.84%	reich Äußerung Anschluß	
29	C-09	68	0.80%	ns Äußerung kritisch	
30	C-13	57	0.67%	Behauptung Göring Jude	
31	C-37	57	0.67%	Umgang verboten kriegsgelangen	
32	C-42	53	0.62%	Amtsunterschlagung Stgb re	
33	C-29	45	0.53%	Sender Abhören ausländisch	
34	C-34	37	0.43%	Körperverletzung gefährlich ö	
35	C-30	36	0.42%	U hg deutsch	
36	C-40	35	0.41%	Fleisch Bezug Mark	
37	C-22	34	0.40%	Erzählung Dachau hg	
38	C-31	32	0.38%	Kritik Jude angeblich	
39	C-32	26	0.30%	Wehrmacht hg Stgb	
40	C-41	26	0.30%	Kofferdiebstahl pferdeverkäufen Preisüberschreitung	
41	C-43	26	0.30%	verfahren beenden Stgb	
42	C-44	26	0.30%	Predigt geistlich hg	
43	C-04	25	0.29%	Hitler hg Attentat	
44	C-23	25	0.29%	Devisenvergehen hg England	
45	C-33	25	0.29%	Bemerkung Brief hg	
46	C-45	25	0.29%	Wahl Führer Behauptung	

Figure 4. Overview of cluster results of the inductive-computational approach with top three keywords for each.

## Comparison of the Different Approaches and Conclusion

As described, the different categorisation approaches offer various and specialised views on the source material. The comparison of the individual case numbers per category (cf. Figure 5 as well as compared to Figure 4) already makes that clear.

## Deductive-(Legal-)Philosophical Approach

Categories	# of cases
D-1 <i>Heimtückegesetz</i>	344
D-2 Reich Penal Code	195
D-3 war related NS Special Penal Law Regulations	583
D-4 Military Penal Code	3
D-5 <i>Heimtückegesetz</i> AND Reich Penal Code	67
D-6 Reich Penal Code AND war related regulations	635
D-7 <i>Heimtückegesetz</i> AND war related regulations	13
D-0 uncategoryed	6691

## Pragmatic-Explorative Approach

Categories	# of cases
E-1 bodily harm and morality	264
E-2 fraud	308
E-3 theft and bribery	909
E-4 war-related economic crimes	644
E-5 prisoners of war related	661
E-6 regime criticism incl. religious and political persecution	238
E-7 conspiracy and protest	528
E-8 negative remarks and hearsay	3005
E-9 person-related remarks	1523
E-0 uncategoryed	451

Figure 5. Case numbers per category for the first two approaches.

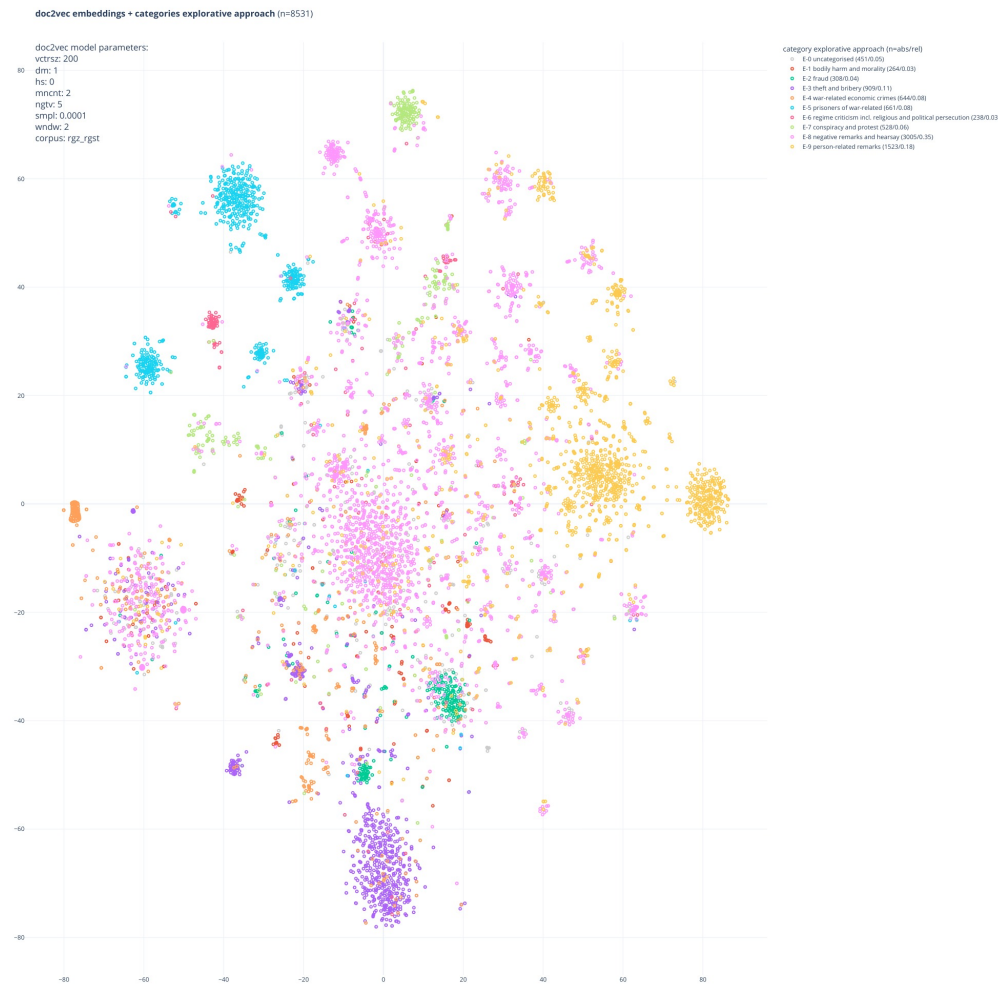
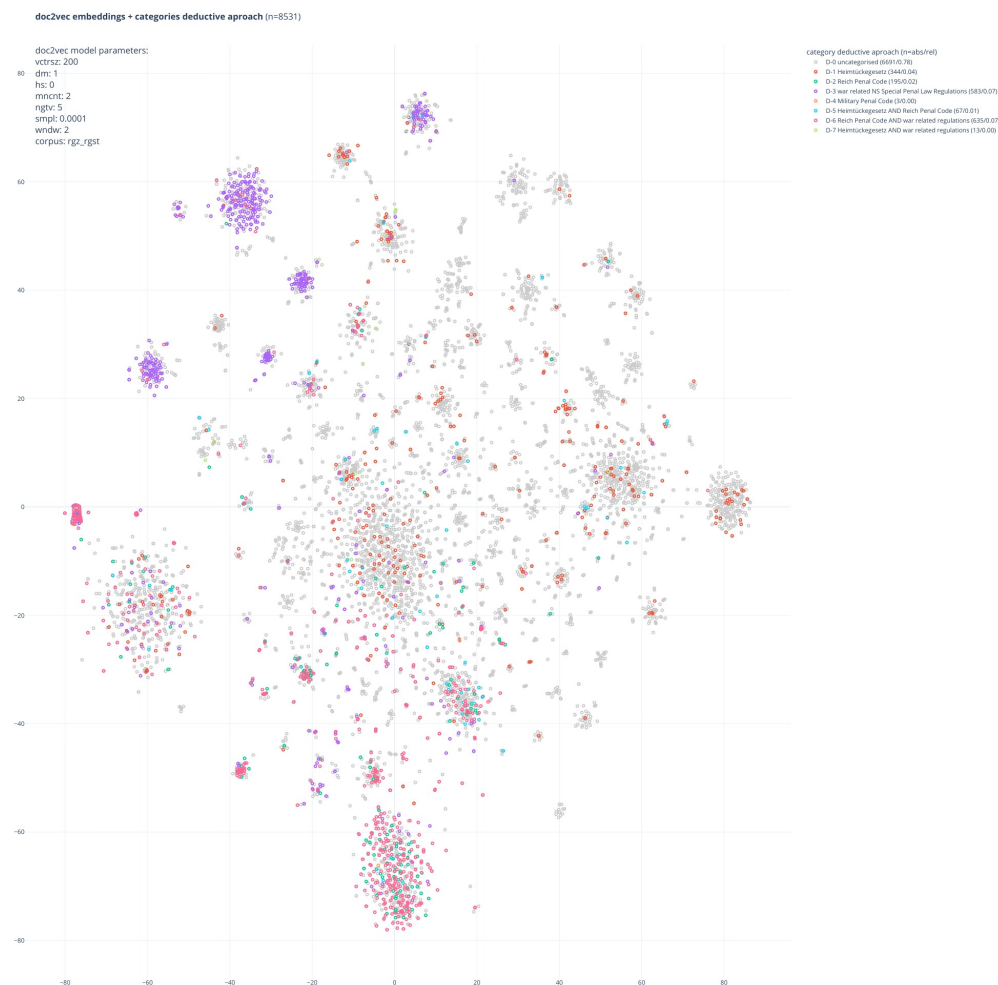


Figure 6. Document embeddings and clusters of the inductive-computational approach with colours/categories from the pragmatic-explorative approach.



**Figure 7.** Figure 6. Document embeddings and clusters of the inductive-computational approach with colours/categories from the pragmatic-explorative approach. Figure 7. Document embeddings and clusters of the inductive-computational approach with colours/categories from the deductive-(legal-)philosophical approach.

Nevertheless, some degree of overlap can be identified as well (cf. Figures 6 and 7). For several of the categories formed in the pragmatic-explorative approach, for example, it can be said that the offences subsumed under each category were criminalised in the first place by a very specific Nazi special criminal law ordinance: 38

The simplest compatibility of one of the pragmatic-explorative categories is in category E-5 prisoners of war-related crimes. An indictment or conviction in this case always took place under §4 of the “*Wehrkraftverordnung*” and is therefore also part of the deductive category D-3 (as can also be seen in the overlaps and distinctive colouring of clusters in the top-left quadrant in Figures 6 and 7). As already mentioned, C-35, C-37, and C-38 combined show a similar composition of cases. Conformity in this instance is predominantly owed to the wording used by the editors in the 1970s and the clearly defined legal basis for prisoners of war-related crimes. 39

Three of these pragmatic-explorative categories even refer to a single Nazi special criminal law ordinance, namely the “*Heimtückegesetz*”, or to be even more precise, to three concrete paragraphs of the “*Heimtückegesetz*” (§§1, 2, and 4), namely E-8 negative remarks and hearsay and E-9 person-related remarks. Offences subsumed in E-6 regime criticism including religious and political persecution, however, were also based on the “Reichstag Fire Decree” in some clearly identifiable cases – always against Jehovah’s Witnesses (known and listed in the records as *Ernstes Bibelforscher*). This shows the qualitatively, but also quantitatively extremely large role that the “*Heimtückegesetz*” had for the juridical practice of the Nazi special courts. Through the inductive-computational approach, we find that this distinction is made even more clear in the category C-28. 40

For “regular” offences where prosecution by the special court was primarily based on very specific paragraphs of regular codified criminal law, similar correspondences between the categorisation approaches can be determined. Overall, for the diachronic analysis of the 1970s “Archive Inventory” regarding the various legal sources used as a basis for the NS Special Court proceedings, this at least partial factual correspondence of the pragmatic-explorative categorisation carried out along semantic terms with certain laws, ordinances, or even certain legal paragraphs helps to gain a better understanding for the source material as well as for a more thorough reflection on what the categories represent. This enables us to choose the most suitable approach for different questions and also to better take into account the necessary theoretical background (cf. above, remarks on categorisation).

41

Through the different categorisation approaches, all of which took into account the various theoretical considerations of categorisation as initially stated in this paper, we now have templates for processing those inquiries. Next steps also include a refinement of the extracted information from the hOCR files in order to maximise the number of valid cases that can be incorporated in the analysis as well as to get a more accurate picture of categorising them. Successful categorisation – not only of charges but of other relevant information such as results of the proceedings – also provides the basis for more far-reaching analyses. For example, it is planned to conduct correspondence analyses on different aspects, like (in-)dependencies between socio-demographic markers and charges as well as results. The workflow initially adopted the terminology used in the 1970s which is essentially based on the source terms and not yet on a controlled and reduced vocabulary. Since statistical/quantitative in-depth analyses are not very fruitful when based on a large number of source terms as they tend to be less significant and more confusing in the (visual) representation, data has to be aggregated employing different kinds of categories, for instance on questions of social status or professions. For this paper, we focused on accusations. Yet, our considerations apply analogously to other categories to be formed. The presented deliberations on categorisation are to be understood in preparation for the upcoming historical investigation of the data with the help of this kind of statistical method.

42

#### Notes

43

[1] Authors are listed in alphabetical order.

44

[2] Before unstitching the books and using an overhead book scanner instead of a document feeder scanner, we were able to extract only 7,778 out of 9,955 inventory records. Eliminating page warping and skew in the digitised images allowed us to extract an additional 753 inventory records and improve our result by 8 %.

45

[3] We would like to thank Andreas Einwiller for programming a Python script extracting JSON records from hOCR-files.

46

[4] A detailed overview of all National Socialist special penal regulations enacted in the years 1933–1944 can be found in: BayStMUK Part 7: Register 1, 1977, pp. 2338–2346 [BayStMUK 1975–1977].

47

[5] *Verordnung über außerordentliche Rundfunkmaßnahmen*, enacted on 1 September 1939; RGBI I 1939, Nr. 169, p. 1681.

48

[6] As an example, the query part for E-1 presents itself as follows: WENN(ODER(ISTZAHL(SUCHEN("sittlich"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Notzucht";[@[text\_normalized]]));ISTZAHL(SUCHEN("Unzucht"; [@[text\_normalized]]));ISTZAHL(SUCHEN("züchtig";[@[text\_normalized]]));ISTZAHL(SUCHEN("Abtreibung"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Rauf";[@[text\_normalized]]));ISTZAHL(SUCHEN("Mord"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Raub";[@[text\_normalized]]));ISTZAHL(SUCHEN("Kind"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Totschlag";[@[text\_normalized]]));ISTZAHL(SUCHEN("Tötung"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Angriff";[@[text\_normalized]]));ISTZAHL(SUCHEN("Körper"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Drohung";[@[text\_normalized]]));ISTZAHL(SUCHEN("Brand"; [@[text\_normalized]]));ISTZAHL(SUCHEN("erpress";[@[text\_normalized]]));ISTZAHL(SUCHEN("vergew"; [@[text\_normalized]]));ISTZAHL(SUCHEN("Nötigung";[@[text\_normalized]]));ISTZAHL(SUCHEN("spreng"; [@[text\_normalized]]));"1 - bodily harm and morality"; [if part/subquery for next category follows]).

49

[7] We would like to thank the “Staatsbibliothek zu Berlin – Preußischer Kulturbesitz” for providing us with a larger

50



corpus of similar language, namely the *Entscheidungen des Reichsgerichts in Zivilsachen* [RGZ 1880–1945] and the *Entscheidungen des Reichsgerichts in Strafsachen* [RGSt 1880–1945]. Available at: <https://staatsbibliothek-berlin.de/emedien-meldungen/rgz-rgst> (Accessed: 02 December 2022).

[8] We selected a model trained on the RGZ/RGSt corpus using gensim's Doc2Vec class supplying the following parameters: vector size=300, epochs=200, dm=1, hs=0, mincount=2, negative=5, sample=1e-4, window=2. 51

[9] Clustering is sensitive to the distribution of data. Without knowing the distribution of data, the quality of clustering can be improved by normalising data using quantiles. Available at: <https://developers.google.com/machine-learning/clustering/prepare-data> (Accessed: 02 December 2022). 52

## Works Cited

- Animal Protection Act 1930** Animal Protection Act. (1930) “Tierschutzgesetz’ vom 24. November 1933” in *RGBl I 1933*, Nr. 132, pp. 987–989.
- Baierer 2020** Baierer, Konstantin. (2020) hOCR – OCR Workflow and Output embedded in HTML. Available at: <http://kba.github.io/hocr-spec/1.2/> (Accessed: 22 November 2022).
- BayStMUK 1975–1977** BayStMUK. (1975–1977) *Widerstand und Verfolgung in Bayern 1933–1945. Hilfsmittel, im Auftrag des Bayerischen Staatsministeriums für Unterricht und Kultus (BayStMUK), herausgegeben von der Generaldirektion der Staatlichen Archive Bayerns. Archivinventare, Bd. 3: Sondergericht München, 7 Teile. München: Eigendruck des Staatsarchivs München / Buchbindung Schmidkonz Regensburg.*
- Bod 2018** Bod, Rens. (2018) “Modelling in the Humanities: Linking Patterns to Principles” in Ciula, Arianna et al. (ed.) *Models and Modelling between Digital and Humanities – A Multidisciplinary Perspective. Historical Social Research Supplement*, 31, pp. 78–95.
- Bracher 1960** Bracher, Karl Dietrich. (1960) “Stufen der Machtergreifung” in Bracher, Karl Dietrich, Sauer, Wolfgang, and Schulz, Gerhard (ed.) *Die nationalsozialistische Machtergreifung. Studien zur Errichtung des totalitären Herrschaftssystems in Deutschland 1933/34*. Köln: Westdeutscher Verlag, pp. 31–368.
- Christians 2020** Christians, Annemone. (2020) *Das Private vor Gericht. Verhandlungen des Eigenen in der nationalsozialistischen Rechtspraxis*. Göttingen: Wallstein Verlag.
- Donig and Rehbein 2022** Donig, Simon and Rehbein, Malte. (2022) “Für eine ‘gemeinsame digitale Zukunft’. Eine kritische Verortung der Digital History” in *Geschichte in Wissenschaft und Unterricht*, 9/10, pp. 527–545.
- Durkheim and Mauss 1901** Durkheim, Emile and Mauss, Marcel. (1901) “DE QUELQUES FORMES PRIMITIVES DE CLASSIFICATION. CONTRIBUTION A L’ÉTUDE DES REPRÉSENTATIONS COLLECTIVES” in *L’Année sociologique*, 6, pp. 1–72.
- Dörner 1998** Dörner, Bernward. (1998) *“Heimtücke”: Das Gesetz als Waffe. Kontrolle, Abschreckung und Verfolgung in Deutschland 1933–1945*, Paderborn / München / Wien / Zürich: Ferdinand Schöningh.
- Ester et al. 1996** Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, and Xu, Xiaowei. (1996) “A density-based algorithm for discovering clusters in large spatial databases with noise” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96)*, pp. 226–231.
- Feuerbach 1801** Feuerbach, Paul Johann Anselm [von]. (1801) *Lehrbuch des gemeinen in Deutschland geltenden Peinlichen Rechts*. Gießen: Georg Friedrich Heyer.
- Fichte 1966/70** Fichte, Johann Gottlieb. (1966/70) “Grundlage des Naturrechts nach Prinzipien der Wissenschaftslehre” [first edition 1796] in Jacob, Hans and Lauth, Reinhard (ed.) *J. G. Fichte-Gesamtausgabe der Bayerischen Akademie der Wissenschaften*. Stuttgart-Bad Canstatt: Friedrich Frommann Verlag (Günther Holzboog), Werke, Vol. 3: pp. 291–460, Vol. 4: pp. 1–166.
- Fröhlich 1983** Fröhlich, Elke. (1983) *Die Herausforderung des Einzelnen. Geschichten über Widerstand und Verfolgung (= Bayern in der NS Zeit. Herausgegeben von Martin Broszat und Elke Fröhlich, VI)*. München / Wien: R. Oldenbourg Verlag.
- Garbe 1999** Garbe, Detlef. (1999) *Zwischen Widerstand und Martyrium. Die Zeugen Jehovas im “Dritten Reich” (= Studien zur Zeitgeschichte, 42)*, 4th edition. München: R. Oldenbourg Verlag.
- Gemoll 2006** Gemoll. (2006) *Griechisch-deutsches Schul- und Handwörterbuch von W. Gemoll und K. Vretska*, 10th

- Gerstmeier et al. 2022** Gerstmeier, Markus, Donig, Simon, Gassner, Sebastian, and Rehbein, Malte. (2022) "Die Archivinventare zum Sondergericht München (1933–1945) digital. Quellenwert – Verdattung – Erkenntnisperspektiven" in *Archivalische Zeitschrift*, 99 (1), pp. 215–251.
- Ginzburg and Poni 1991** Ginzburg, Carlo and Poni, Carlo. (1991) "The Name and the Game. Unequal Exchange and the Historiographic Marketplace" in Muir, Edward Wallace and Ruggiero, Guido (ed.) *Microhistory and the Lost Peoples of Europe*. Selections from Quaderni storici. Baltimore: MD, pp. 1–10. Available at: <http://opac.regesta-imperii.de/id/2114403> (Accessed: 22 November 2022).
- Graczyk 2021** Graczyk, Konrad. (2021) *Ein anderes Gericht in Oberschlesien. Sondergericht Kattowitz 1939–1945* (= Beiträge zur Rechtsgeschichte des 20. Jahrhunderts, 119). Tübingen: Mohr Siebeck.
- Graham, Milligan, and Weingart 2016** Graham, Shawn, Milligan, Ian, and Weingart, Scott B. (2016) *Exploring big historical data. The historian's macroscope*. London: Imperial College Press. Available at: <http://www.themacroscope.org/2.0/> (Accessed: 22 November 2022).
- Gruchmann 2001** Gruchmann, Lothar. (2001) *Justiz im Dritten Reich 1933–1940. Anpassung und Unterwerfung in der Ära Gürtner* (= Quellen und Darstellungen zur Zeitgeschichte, 28), 3rd edition. München: R. Oldenbourg Verlag.
- Hackl 2022** Hackl, Gabriele. (2022) "Über das Mögliche urteilen. Urteilsfindung in Kriegswirtschaftsverfahren vor dem Sondergericht Wien" in Ruby, Sigrid and Krause, Anja (ed.) *Sicherheit und Differenz in historischer Perspektive / Security and Difference in Historical Perspective*. Baden-Baden: Nomos Verlag, pp. 241–259.
- Harris 2020** Harris, Charles R. et al. (2020) "Array programming with NumPy", *Nature*, 585(7825), pp. 357–362. Available at <https://doi.org/10.1038/s41586-020-2649-2> (Accessed: 03 December 2020).
- Heimtückegesetz 1934** Heimtückegesetz. (1934) "Gesetz gegen heimtückische Angriffe auf Staat und Partei und zum Schutz der Parteiuniformen, 29 December 1934" in *RGBl I 1933*, Nr. 137, pp. 1269–1271, preceded by the "Verordnung des Reichspräsidenten zur Abwehr heimtückischer Angriffe gegen die Regierung der nationalen Erhebung, 21 March 1933" in *RGBl I 1933*, Nr. 24, p. 135.
- Hensle 2005** Hensle, Michael. (2005) *"Rundfunkverbrechen" vor nationalsozialistischen Sondergerichten. Eine vergleichende Untersuchung der Urteilspraxis in der Reichshauptstadt Berlin und der südbadischen Provinz*. Von der Fakultät I Geisteswissenschaften der Technischen Universität Berlin genehmigte Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (2001). Available at: DOI 10.14279/depositonce-1208 (Accessed: 22 November 2022).
- Holland et al. 1986** Holland, John H., Holyoak, Keith J., Nisbett, Richard E., and Thagard, Paul R. (1986) *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, Massachusetts: MIT Press.
- Hüttenberger 1981** Hüttenberger, Peter. (1981) "Heimtückefälle vor dem Sondergericht München 1933–1939" in Broszat, Martin, Fröhlich, Elke, and Grossmann, Anton (ed.) *Bayern in der NS-Zeit IV: Herrschaft und Gesellschaft im Konflikt, Teil C*. München / Wien: R. Oldenbourg Verlag, pp. 435–526.
- Kant 1787/1904** Kant, Immanuel. (1787/1904) "Kritik der reinen Vernunft. Zweite, hin und wieder verbefterte Auflage" in Preußische Akademie der Wissenschaften (ed.) *Kant's gesammelte Schriften, Band III*. Berlin: Reimer.
- Kelsen 2017** Kelsen, Hans. (2017) *Reine Rechtslehre* [first edition: Leipzig / Wien 1934]. Mit einem Anhang: Das Problem der Gerechtigkeit. Studienausgabe der 2. Auflage 1960, ed. by Matthias Jestaedt. Tübingen: Mohr Siebeck.
- Klemperer 1947/2020** Klemperer, Victor. (1947/2020) *LTI. Tagebuch eines Philologen*. Herausgegeben von Elke Fröhlich. Stuttgart: Philipp Reclam Jun. [first edition: Berlin: Aufbau-Verlag]
- Lahusen 2022** Lahusen, Benjamin. (2022) *Der Dienstbetrieb ist nicht gestört. Die Deutschen und ihre Justiz 1943–1948*. München: C. H. Beck.
- Lau and Baldwin 2016** Lau, Jey Han and Baldwin, Timothy. (2016) "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation". arXiv. Available at: <https://doi.org/10.48550/arXiv.1607.05368> (Accessed: 02 December 2022).
- Le and Mikolov 2014** Le, Quoc V. and Mikolov, Tomas. (2014) "Distributed Representations of Sentences and Documents". arXiv. Available at: <https://arxiv.org/abs/1405.4053v2> (Accessed: 02 December 2022).
- MStGB 1943** MStGB. (1943) "Strafrecht der deutschen Wehrmacht: Militärstrafgesetzbuch, Kriegssonderstrafrechtsverordnung, Kriegsstrafverfahrensordnung, Wehrmachtdisziplinarstrafordnung, Beschwerdeordnung, Sondergerichtsbarkeit für Angehörige der SS und Polizeiverbände, Reichsstrafgesetzbuch und

zahlreiche andere Bestimmungen". München: C. H. Beck.

- McKinney 2010** McKinney, Wes. (2010) "Data Structures for Statistical Computing in Python" in *Proceedings of the 9th Python in Science Conference*, pp. 56–61. Available at: DOI 10.25080/Majora-92bf1922-00a (Accessed: 02 December 2022).
- Oehler 1997** Oehler, Christiane. (1997) *Die Rechtsprechung des Sondergerichts Mannheim 1933–1945* (= Freiburger Rechtsgeschichtliche Abhandlungen. Neue Folge, 25). Berlin: Duncker and Humblot.
- Ordinance on War Economy 1939** Ordinance on War Economy. (1939) "'Kriegswirtschaftsverordnung' vom 4. September 1939" in *RGBl I 1939*, Nr. 163, pp. 1609–1613 and 1700.
- Pedregosa 2011** Pedregosa, F. et al. (2011) "Scikit-learn: Machine Learning in Python" in *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Penal Ordinance against Polish and Jews 1941** Penal Ordinance against Polish and Jews. (1941) "'Verordnung über die Strafrechtspflege gegen Polen und Juden in den eingegliederten Ostgebieten' vom 4. Dezember 1941" in *RGBl I 1941*, Nr. 140, pp. 759–761.
- RGBl I 1922–1945** RGBl I. (1922–1945) *Deutsches Reichsgesetzblatt, Teil I, herausgegeben vom Reichsministerium des Innern*. Berlin: Reichsverlagsamt.
- RGSt 1880–1945** RGSt. (1880–1945) *Entscheidungen des Reichsgerichts in Strafsachen. Herausgegeben von den Mitgliedern des Gerichtshofes und der Reichsanwaltschaft*, 77 Volumes. Berlin – Leipzig: Verlag Veit und Comp. (Vol. 78, concerning 1945's decisions, was published in 2007: Berlin – Boston: Verlag Walter de Gruyter). Available at: <https://rgst-1staatsbibliothek-2berlin-1de-10099dfd10212.erf.sbb.spk-berlin.de/> (Accessed: 02 December 2022).
- RGZ 1880–1945** RGZ. (1880–1945) *Entscheidungen des Reichsgerichts in Zivilsachen 1880 bis 1945. Herausgegeben von den Mitgliedern des Gerichtshofes und der Reichsanwaltschaft*, 173 Volumes. Berlin – Leipzig: Verlag Veit und Comp. Available at: <https://rgz-1staatsbibliothek-2berlin-1de-10099dfd10212.erf.sbb.spk-berlin.de/> (Accessed: 02 December 2022).
- RStGB 1871, 1935** RStGB. (1871, 1935) *Strafgesetzbuch für das Deutsche Reich vom 15. Mai 1871* – appeared until 1945 in different editions, e. g. 1935: *Strafgesetzbuch für das Deutsche Reich. Mit Erläuterungen und einem Anhang, enthaltend strafrechtliche Nebengesetze und Notverordnungen. Begründet von Julius von Staudinger. Neubearbeitet von Hermann Schmitt*, 20th edition. München / Berlin: C. H. Beck.
- Rehurek and Petr 2010** Rehurek, Radim and Sojka Petr. (2010) "Software Framework for Topic Modelling with Large Corpora" in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta: ELRA, pp. 45–50. Available at: <https://radimrehurek.com/gensim/index.html> (Accessed: 02 December 2022).
- Reichstag Fire Decree 1933** Reichstag Fire Decree. (1933) "'Verordnung des Reichspräsidenten zum Schutz von Volk und Staat' vom 28. Februar 1933" in *RGBl I 1933*, p. 83.
- Reul et al. 2018** Reul, Christian et al. (2018) "State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines". arXiv. Available at: <http://arxiv.org/abs/1810.03436> (Accessed: 15 November 2022).
- Richter 2021** Richter, Hedwig. (2021) *Demokratie. Eine deutsche Affäre. Vom 18. Jahrhundert bis zur Gegenwart*. Bonn: Bundeszentrale für politische Bildung.
- Schlumbohm and Gribaudo 1998** Schlumbohm, Jürgen and Gribaudo, Maurizio. (ed.) (1998) *Mikrogeschichte – Makrogeschichte. Komplementär oder inkommensurabel? Göttinger Gespräch zur Geschichtswissenschaft* (= Göttinger Gespräche zur Geschichtswissenschaft, 7). Göttingen: Wallstein-Verl.
- Schmitt 1922/2021** Schmitt, Carl. (1922/2021) *Politische Theologie. Vier Kapitel zur Lehre von der Souveränität* [first edition 1922]. 11th edition. Berlin: Duncker und Humblot.
- Seiffert 2006** Seiffert, Helmut. (2006) *Einführung in die Wissenschaftstheorie. Zweiter Band: Geisteswissenschaftliche Methoden: Phänomenologie – Hermeneutik und historische Methode – Dialektik*. 11th edition. München: C. H. Beck.
- Slaughter Tax Act 1930** Slaughter Tax Act. (1930) "'Gesetz zur Abgleichung des ordentlichen Staatshaushalts (Nr. 48110)' vom 31. Oktober 1930" in *Gesetz- und Verordnungs-Blatt für den Freistaat Bayern*, Nr. 34, München, 3. November 1930, pp. 327–332.
- Smith 2007** Smith, Ray. (2007) "An Overview of the Tesseract OCR Engine" in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2. Curitiba, Parana, Brazil: IEEE, pp. 629–633. Available at: <http://ieeexplore.ieee.org/document/4376991/> (Accessed: 15 November 2022).

**TLL 1906–1912** TLL. (1906–1912) *Thesaurus linguae Latinae, Vol. III: C – Comvs, ed. auctoritate et consilio academiarm qvinqve Germanicarvm: Berolinensis, Gottingensis, Lipsiensis, Monacensis, Vindobonensis*. Lipsiae: In aedibus B. G. Teubneri.

**Verordnung 1933** Verordnung. (1933) “Verordnung der Reichsregierung über die Bildung von Sondergerichten’ vom 21. März 1933” in *RGBl I 1933*, Nr. 24, pp. 136–138.

**Volksschädlingsverordnung 1939** Volksschädlingsverordnung. (1939) “Verordnung gegen Volksschädlinge’ vom 5. September 1939” in *RGBl 1939 I*, Nr. 168, p. 1679.

**Weber and Piazzolo 1998** Weber, Jürgen and Piazzolo, Michael. (1998) “Parteisoldaten in Richterrobe” in Weber, Jürgen and Piazzolo, Michael (ed.) *Justiz im Zwielficht. Ihre Rolle in Diktaturen und die Antwort des Rechtsstaates*. München: Olzog, pp. 11–22.

**Wehler 1972** Wehler, Hans-Ulrich. (ed.) (1972) *Geschichte und Soziologie*. Köln: Kiepenheuer und Witsch.

**Wehrkraftverordnung 1939** Wehrkraftverordnung. (1939) “Verordnung zur Ergänzung der Strafvorschriften zum Schutze der Wehrkraft des Deutschen Volkes’ vom 25. November 1939” in *RGBl I 1939*, Nr. 238, p. 3219.

**Wick, Reul, and Puppe 2018** Wick, Christoph, Reul, Christian, and Puppe, Frank. (2018) “Calamari. A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition”. arXiv. Available at: <http://arxiv.org/abs/1807.02004> (Accessed: 15 November 2022).

**Wildt 2007** Wildt, Michael. (2007) “Das ‘Bayern-Projekt’, die Alltagsforschung und die ‘Volksgemeinschaft’” in Frei, Norbert (ed.) *Martin Broszat, der “Staat Hitlers” und die Historisierung des Nationalsozialismus*, Göttingen: Wallstein Verlag, pp. 119–129.

**Wildt 2019** Wildt, Michael. (2019) *Die Ambivalenz des Volkes. Der Nationalsozialismus als Gesellschaftsgeschichte*, 2nd edition. Berlin: Suhrkamp.

**Wogersien 2007** Wogersien, Maik. (2007) “Allgemeines ‘unpolitisches Strafrecht’ als Kriegsstrafrecht vor den Sondergerichten” in Daubach, Helia-Verena (ed.) “... eifrigster Diener und Schützer des Rechts, des nationalsozialistischen Rechts ...”. *Nationalsozialistische Sondergerichtsbarkeit. Ein Tagungsband* (= Juristische Zeitgeschichte Nordrhein-Westfalen, 15). Düsseldorf: Eigenverlag des Justizministeriums des Landes Nordrhein-Westfalen, pp. 63–72.

**van der Maaten and Hinton 2008** van der Maaten, L. J. P. and Hinton, G. E. (2008) “Visualizing High-Dimensional Data Using t-SNE” in *Journal of Machine Learning Research*, 9, pp. 2579–2605.



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