Hefte zur Bildungs- und Hochschulforschung

Franziska Weeber, Thomas Hinz, Jasmin Meyer and Frank Multrus

Rule-based semi-automated coding procedure to classify fields of study in surveys among university students in Germany

Universität Konstanz; Arbeitsgruppe Hochschulforschung

Hefte zur Bildungs- und Hochschulforschung (99)

Februar 2024

Higher Education Research Group (AG Hochschulforschung),

University of Konstanz,

February 2024

https://www.soziologie.uni-konstanz.de/en/research-group-on-higher-education/

ISSN 1616-0398

DOI: https://doi.org/10.48787/kops/352-2-15qrl6ayxtl7u8

Cite as: Weeber, Franziska; Hinz, Thomas; Meyer, Jasmin; Multrus, Frank, 2024. Rule-based semi-automated coding procedure to classify fields of study in surveys among university students in Germany. Hefte für Bildungs- und Hochschulforschung 99. University of Konstanz. DOI: 10.48787/kops/352-2-15qrl6ayxtl7u8

Contents

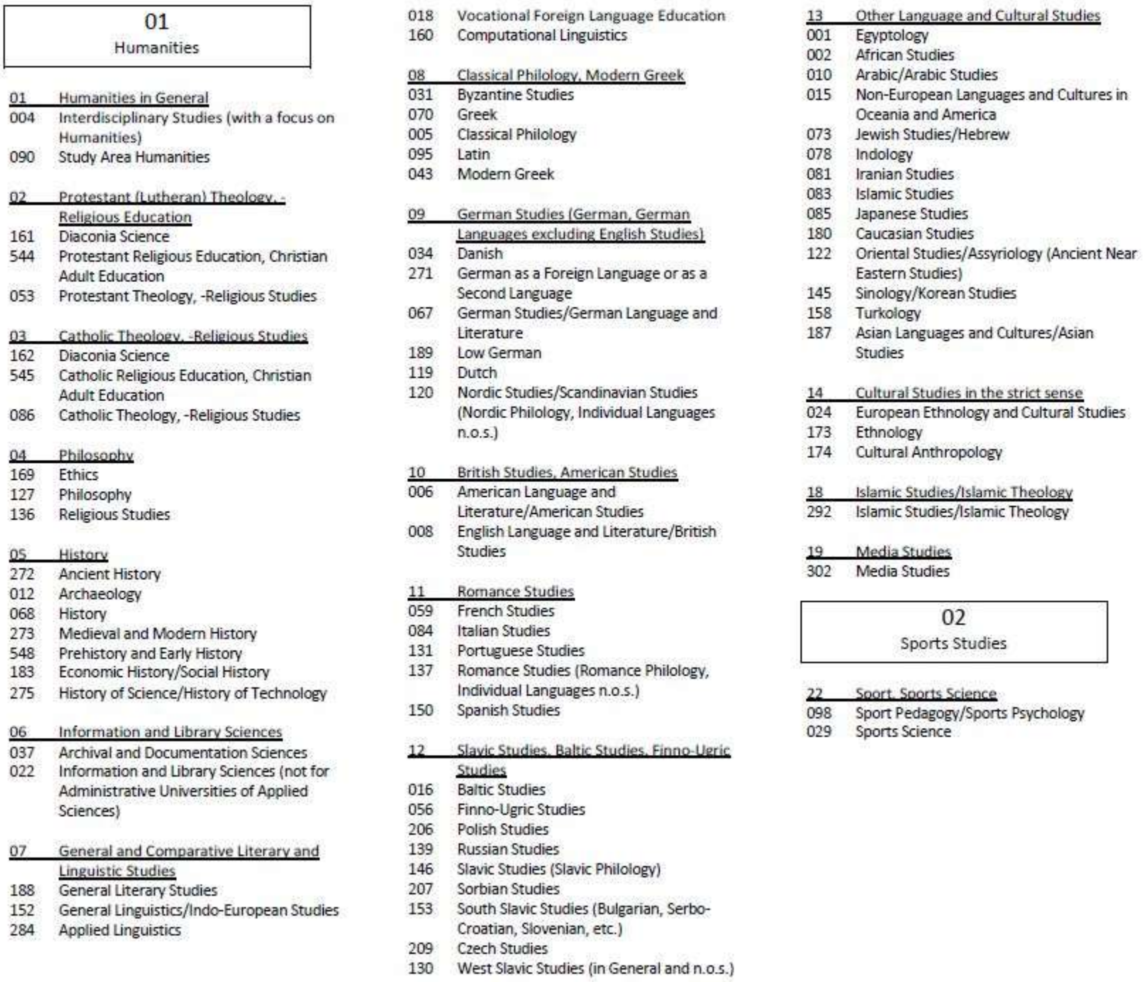
Introduction

Converting natural language written responses to a code from standardized category system typically is a costly task regarding time and expertise. In previous large-scale surveys, the classification into a detailed and frequently used category system has been done manually. For a large Germany wide survey among university students from 250 universities ("Die Studierendenbefragung in Deutschland", SiD, Beuße et al. 2022), we developed a procedure that enables semi-automated classification of fields of study that respondents reported in natural language. Our procedure combines use-case adapted preprocessing with a variety of rule-based machine learning tasks. In the end, the application yields a large reduction of manual effort: While the original study contained around 324,000 responses to the fields of study questions, our method is able to code most of them automatically and provide suggestions that need manual checking for 13,435 values (around 4.1 percent of all responses). For only 2,359 values (around 0.73 percent of all responses), the tool cannot produce an automated coding or a suggestion and requires a manual coding. We also review other approaches to (semi-)automated classification and explain why we chose the approach. We then describe the procedure in detail while simultaneously explaining how it can easily be adapted to other use cases.

Problem statement

As a starting point to develop the classifier, we are faced with a huge number of open-ended answers containing study programs from a large-scale survey among university students (SiD, conducted in summer 2021). The total number of open-ended answers is around 324,000. In order to get a fine-grained and reliable measure of the field of study (i.e. a key variable for many substantive analyses), the overall goal of the coding procedure is to assign the correct code from the Statistisches Bundesamt (Federal Statistical Office, also: Destatis) study subject classification to each answer. This classification scheme, in the following *Destatis classification scheme* consists of 273 fields of study which are hierarchically structured (with general classification containing more fine-grained subcategories, see Figure 1; German version in the appendix Figure 1A). The assignment of open-ended text to these 273 codes should be achieved as efficiently as possible without sacrificing the coding quality, i.e. reducing wrong category assignment (false positives).

Figure 1: Field of study classification (Destatis classification scheme from Statistisches Bundesamt, example: humanities)



01
Humanities

01 Humanities in General
004 Interdisciplinary Studies (with a focus on Humanities)
090 Study Area Humanities

02 Protestant (Lutheran) Theology, -Religious Education
161 Diaconia Science
544 Protestant Religious Education, Christian Adult Education
053 Protestant Theology, -Religious Studies

03 Catholic Theology, -Religious Studies
162 Diaconia Science
545 Catholic Religious Education, Christian Adult Education
086 Catholic Theology, -Religious Studies

04 Philosophy
169 Ethics
127 Philosophy
136 Religious Studies

05 History
272 Ancient History
012 Archaeology
068 History
273 Medieval and Modern History
548 Prehistory and Early History
183 Economic History/Social History
275 History of Science/History of Technology

06 Information and Library Sciences
037 Archival and Documentation Sciences
022 Information and Library Sciences (not for Administrative Universities of Applied Sciences)

07 General and Comparative Literary and Linguistic Studies
188 General Literary Studies
152 General Linguistics/Indo-European Studies
284 Applied Linguistics
018 Vocational Foreign Language Education
160 Computational Linguistics

08 Classical Philology, Modern Greek
031 Byzantine Studies
070 Greek
005 Classical Philology
095 Latin
043 Modern Greek

09 German Studies (German, German Languages excluding English Studies)
034 Danish
271 German as a Foreign Language or as a Second Language
067 German Studies/German Language and Literature
189 Low German
119 Dutch
120 Nordic Studies/Scandinavian Studies (Nordic Philology, Individual Languages n.o.s.)

10 British Studies, American Studies
006 American Language and Literature/American Studies
008 English Language and Literature/British Studies

11 Romance Studies
059 French Studies
084 Italian Studies
131 Portuguese Studies
137 Romance Studies (Romance Philology, Individual Languages n.o.s.)
150 Spanish Studies

12 Slavic Studies, Baltic Studies, Finno-Ugric Studies
016 Baltic Studies
056 Finno-Ugric Studies
206 Polish Studies
139 Russian Studies
146 Slavic Studies (Slavic Philology)
207 Sorbian Studies
153 South Slavic Studies (Bulgarian, Serbo-Croatian, Slovenian, etc.)
209 Czech Studies
130 West Slavic Studies (in General and n.o.s.)

13 Other Language and Cultural Studies
001 Egyptology
002 African Studies
010 Arabic/Arabic Studies
015 Non-European Languages and Cultures in Oceania and America
073 Jewish Studies/Hebrew
078 Indology
081 Iranian Studies
083 Islamic Studies
085 Japanese Studies
180 Caucasian Studies
122 Oriental Studies/Assyriology (Ancient Near Eastern Studies)
145 Sinology/Korean Studies
158 Turkology
187 Asian Languages and Cultures/Asian Studies

14 Cultural Studies in the strict sense
024 European Ethnology and Cultural Studies
173 Ethnology
174 Cultural Anthropology

18 Islamic Studies/Islamic Theology
292 Islamic Studies/Islamic Theology

19 Media Studies
302 Media Studies

02
Sports Studies

22 Sport, Sports Science
098 Sport Pedagogy/Sports Psychology
029 Sports Science

Source: Statistisches Bundesamt 2023, German version in the appendix (Figure 1A)

In previous surveys with a significantly lower number of cases, human coders have manually assigned the correct code to the written answers. While this method requires plenty of time and expertise, one can expect a high classification accuracy which is important for the further use of this key information on fields of study in further data analysis. Using the work of manual coders, we still do not expect a perfect classification accuracy of 100 percent correct assignments. False assignments can be the result of different assignments of edge cases between coders, i.e., a low inter-coder reliability (Krippendorff 1978), inconsistencies within the coding of single coders, or simple mistakes. The semi-automated procedure should at least reach a comparably low error rate to the best practice manual coding procedure.

Data

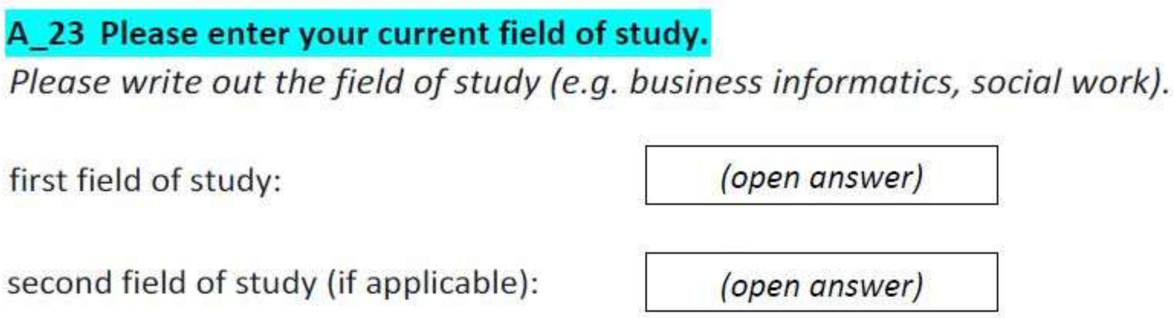
Data input that is to be coded stems from an online survey among university students in Germany in 2021 (SiD). We use open-ended responses to questions on current and previous study programs in which students are or have been enrolled (Beuße et al. 2022). The seemingly simple and straightforward question about the field of study is not easily framed in student surveys by means of predefined, distinct categories. This is firstly because the existing BA and MA programs incorporate content from various academic disciplines. Particularly, BA programs often combine major and minor subjects. Secondly, when students are asked about their field of study, they often think of the specific name of the study program at their university (e.g. “LKM”: Literatur, Kunst, Medien; literature, arts, and media), from which the field of study may not always be clearly discernable.

Most importantly, in *cross-university* surveys of students – as in the SiD – the recording of fields of study is particularly challenging when it comes to the question format of predefined categories. The range of study programs offered by German universities has become enormously differentiated with the introduction of bachelor’s and master’s degrees. There is an increasing number of MA programs, some are very specialized, some are taught in English and some are interdisciplinary. In addition, the survey includes all study programs for teachers’ education as well. There are manifold variations among universities that offer study programs in teachers’ education, such as the number of subjects required for studies.

Even if an accurate collection of (distinct and complete) answers to closed questions was possible in principle in online surveys (by listing all study programs in extensive drop-down menus), a sufficiently precise, summarized classification into superimposed and predefined categories of subjects, into which respondents can classify themselves, appears extremely challenging to present on a computer screen and impossible on a mobile device. In addition, the assignment of the cognitive representation of respondents (often simply the name of the study program for which they enrolled) to the 273 predefined subjects as listed in the Destatis classification scheme (Figure 1, Statistisches Bundesamt 2023) would require an enormous effort from the respondents. Thus, the question on the field of study is a legitimate candidate for collecting data by respondents filling in natural language information (open-ended question). To mitigate the substantial cognitive effort of having students categorize their study subject themselves, the primary researchers of the SiD study deliberately opted for an open-ended approach in querying the fields of study.

Figure 2 depicts a screenshot of the question on the current field of study (German version in the appendix, Figure 2A). There are two open-ended fields of input. The second field is optional. For subgroups (identified by filter questions), this question format is applied for (up to three) previous enrollments. In total, there are eight open-ended fields containing possible natural language input on the field of study.

Figure 2: Open-ended question on the field of study



A_23 Please enter your current field of study.

Please write out the field of study (e.g. business informatics, social work).

first field of study: *(open answer)*

second field of study (if applicable): *(open answer)*

Source: SiD, Beuße et al. 2022, German version in the appendix (Figure 2A)

In the following, all input data to questions on the current and previous fields of study form the source data D1.

Data issues and potential sources of error

Ideally, each response would contain exactly one correctly spelled field of study. However, there are some common issues across respondents that must be solved when applying the classification method. As already pointed to, respondents might just write in the name of the study program (e.g. “LKM”: Literatur, Kunst, Medien; literature, arts, and media) that not necessarily corresponds to a field of study.

More trivially, some respondents include typos and spelling errors in their responses. Some spelling errors cannot easily be resolved since study subjects can be read similarly to each other. For example, if someone included a typo in *Biology* and accidentally wrote *Giology*, converting *Giology* to *Biology* would have the same Levenshtein distance when using equal weights as *Giology* to *Geology*, thus we might change some values to a different subject than intended. Our method must thus be able to handle spelling errors in some way.

Furthermore, a subgroup of respondents writes an entire sentence instead of just their field of study (e.g. “Currently I am studying computer science.”). Some others also list the degree they are pursuing in their program (e.g., Bachelor of Science, Master of Education) or provide additional context. Additionally, there might be responses which do not contain the study program at all (e.g. “Lehramt”, i.e., heading for a teaching degree; “I do not want to answer”). Moreover, some respondents might list more than one subject per field. This may be due to a misunderstanding of the question, but in some cases respondents are enrolled for three subjects (e.g., when heading for a teaching degree “Lehramt”), while the questions on field of studies only allow two subjects (see Figure 2).

The answers partly consist of English words as there are also study programs taught in English and/or with an English name. We therefore need a method that takes English words into account and codes them properly. All these characteristics of the available data need to be addressed for automated classification.

Additional data

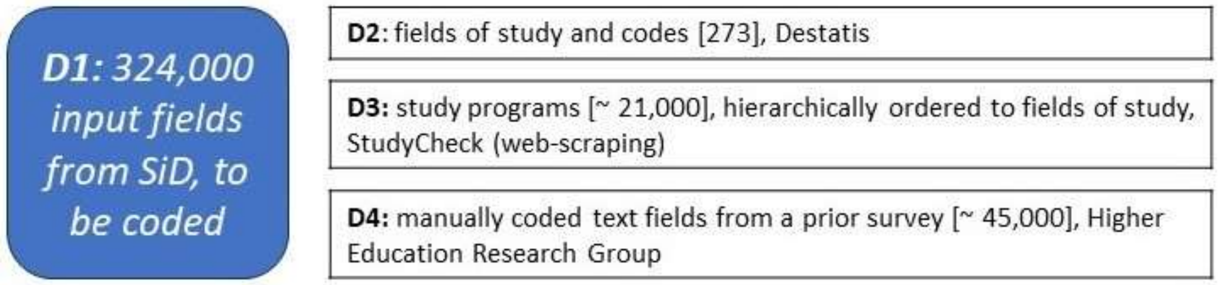
As almost all supervised classifiers, our method requires training data, i.e., data in the same or a similar format the model can use. In practice, we connect existing data on study program names and fields of study (not necessarily given by respondents) with their corresponding code. Concretely, we use three additional data sources as training data and transform them to dictionaries mapping study program names and fields of study to their codes. The listing and naming of fields of study with codes from the Federal Statistical Office (D2; Statistisches Bundesamt), a list of German study programs (D3), and a processed and partially coded data set (D4) from a preliminary survey “Studying in Times of the Corona Pandemic” (Lörz et al. 2020).

- D2: This source file contains the names of the fields of study and the assigned numerical codes from the Federal Statistical Office (Statistisches Bundesamt 2023, see Figure 1). The complete list forms the dictionary to be used further and can be employed to determine direct mapping of plain text data.

- D3: This source file contains an (unofficial) list of all study programs in Germany (around 21,000 programs in October 2023) and was indexed using the website StudyCheck (https://www.studycheck.de/). Based on the search over all categories, subcategories, and fields of study, the latter were retrieved from the website by means of web-scraping. This list is not coded, but in many cases the subcategories correspond to the categories used by the Federal Statistical Office (Statistisches Bundesamt 2023) for fields of study (D2), which allows an assignment of specialized study programs to (higher-level) codes in further cases.
- D4: This component for the development of the classification system contains a partially coded data set obtained from a preliminary survey of university students in the summer of 2020. Manual coding was performed and checked for 45,298 observations. This source file contains the following information: the same plain-text information (as in the current survey) and the processed version, in which the plain text was manually cleaned, and the Destatis classification code (Statistisches Bundesamt 2023) was assigned. This source file can significantly support the classification procedure. D4 dictionary is finally supplemented by a first manual classification of data from the current survey from a preliminary data set from July 2021.

Figure 3 summarizes the four data sources that are used in the procedure.

Figure 3: Data sources (D1, D2, D3, D4) used for classification



D1: 324,000 input fields from SiD, to be coded

D2: fields of study and codes [273], Destatis

D3: study programs [~ 21,000], hierarchically ordered to fields of study, StudyCheck (web-scraping)

D4: manually coded text fields from a prior survey [~ 45,000], Higher Education Research Group

Source: Higher Education Research Group, University of Konstanz

Reasons for relying on rule-based assignments

With the advancement of natural language processing (NLP) techniques and increasing access to large amounts of textual data, more and more computational social science solutions for unstructured textual data have emerged, many of them also applicable to the analysis of open-ended survey questions, in particular embedding models, which transform text data into numeric vectors that maintain text features such as semantic similarity. We will briefly describe the most common methods and our reasoning for relying on rule-based assignments instead of language models. The methods considered here include topic modeling, traditional machine learning classifiers, and state-of-the-art deep learning solutions with large language models (LLMs).

(1) One technique that has been commonly employed in social-science research are topic models (Eshima et al. 2023), an approach extracting word or phrase clusters characterizing underlying aspects of a set of documents, i.e., the documents' topics (Blei et al. 2003; Blei 2012). However, traditional topic modeling cannot take classes defined by the researchers into account,

researchers have to link their classes to topics themselves. Some further developed approaches use researcher-defined topic keywords as the starting point for the topic model (Eshima et al. 2023; Harandizadeh et al. 2022). But even then, it is not guaranteed that the identified topics and classes are similar enough to justify a classification based on a found topic, and it becomes less likely the more classes researchers aim to correctly identify (Pietsch and Lessmann 2018). We did not choose topic models for a simple reason: We have to stick to the Destatis classification scheme with 273 classes.

(2) Card and Smith (2015) evaluate machine learning classifiers, more specifically logistic regression, and recurrent neural networks (RNN), to automatically assign all applicable labels to a survey response (multi-label setting). Using Bayesian optimization to find the best parameter setting for both models, they find the more traditional logistic regression to outperform their RNN approaches. It is also important to note that the highest performance is reached on questions with usually only one label present in the answer and comparatively few possible labels. While Card and Smith (2015) highlight the higher consistency compared of both models to human coders and higher label confidences when using models providing class probabilities, they point to the problem of lacking interpretability for RNN. In comparison to logistic regression or other solutions, the interpretation of its classification decisions is challenging. Because of the somewhat opaque interpretation of classification choices, we decided not to apply these models.

(3) Most state-of-the-art text classifiers in computer science rely on deep language models, such as BERT (Vaswani et al. 2017; Devlin et al. 2019; Liu et al. 2019). These models learn to represent an input sequence of text as a numeric vector (i.e., word embedding) such that each token (i.e., subunits such as words or word pieces) has its own representation that results from its contextual use. This means that the same word with two meanings will also be represented differently. Deploying these models has resulted in new state-of-the-art performances on many NLP tasks including text classification (Devlin et al. 2019). The omnipresence of transformer-based pretrained language models suggests they could be helpful in our project as well, however, we decided not to use them for the following reasons:

- Spelling Errors: Unresolved spelling errors result in many study subjects that the model cannot recognize from the text it was pretrained on, and which are represented differently when converted into numerical input than their correctly spelled counterparts.

- Classification Task: Discrete classification tasks can be divided into multi class (a) and multi label (b) classification tasks. In (a), each response will receive exactly one out of all *k* available labels. In (b), each response can receive any number of labels between 0 and *k*, i.e., all that apply. Our question on the field of study was designed such that we have a multi-class task (a), meaning each study subject should be written in a separate field. However, we have some respondents who entered more than one study program into one response field. We thus have a single-label task in theory, but in practice this would give us multiple incomplete assignments.

- Number of classes: The Destatis classification scheme has 273 classes. Some of these classes occur only rarely while others appear frequently. The large number of classes in combination with the imbalance typically makes the classification for the model much more difficult than having only few and balanced classes (Padurariu and Breaban 2019).

- Identification of false positives: Applying a classifier based on BERT or one of its successors will (in our case) assign one category from the category system to each natural language value. This value is chosen from the class with the maximum output probability. However, one of the main goals of the procedure is to avoid faulty category assignments. To assess the performance of the classifier, one would need a (manually coded) test set. It is possible to identify these cases after prediction by inspecting label probabilities. However, classification models based on transformers have been shown to be overconfident and thus result in less interpretable output probabilities (Schröder and Niekler 2020).

- Missing context: As described, transformers generate a numerical representation of natural language by also considering each token's context. Since we only asked for the study subject, we do not have any context by default. One option would be to include each response value in a template sentence such as "I am studying _____". Although this method might be effective for some values, a few respondents provided additional context or composed complete sentences, which could cause incorrect grammar when using such a template.

Rule-based semi-automated classification procedure

After briefly discussing other, more complex methods and reasoning why they are not suitable for the data at hand, we propose a simple and easily adaptable semi-automated classification procedure including preprocessing, a combination of automatic category assignments, manual category assignments based on suggestions, and purely manual assignments. Our method starts with preprocessing the text to remove as many inconsistencies across entries of the same study program as possible. Then, it aims to identify perfect and almost-perfect matches in the reference dictionaries. All responses that our method can clearly link to a single category will be coded automatically. In cases where the method is uncertain, the response will be presented to a human coder who can choose from a list of category suggestions generated by the coding mechanism. If the method fails to suggest a category, the response will be coded manually without any suggestion. Using this three-step procedure, we avoid generating false-positives during the automated coding which are difficult to identify.

Data preparation (preprocessing)

In a first step, the open data from the survey D1 and the subject names from the source files D2, D3, and D4 are processed in order to identify matches despite minor format deviations. For this purpose, upper case letters are converted to lower case letters, punctuation marks and numbers are removed and replaced by spaces. Double, leading, or trailing spaces are removed. German alphabetic special characters (ä, ö, ü, ß) are converted (ae, oe, ue, ss). Then, any degrees specified are removed from the subjects in D1. To do this, a list of possible labels for degrees and potentially following words is created from the known classified statements and removed from the statement using regular expressions from the specification. Frequently, the term "Lehramt" (or an equivalent title for a teaching degree "Lehramt Grundschule", "Grundschullehramt", i.e. "Primary School Teaching") was given as a subject of study with or without further subject designations. However, in the Destatis classification, "Lehramt" is not categorized as a field of study but as a degree (Bachelor's or Master's of Education or "Staatsexamen", i.e. State Examination). If the term "Lehramt" is mentioned in the clear text data on subjects of study, we define an additional variable to indicate the mention of a teaching degree. Separate Destatis codes are assigned for primary school or special education teacher training. For all other teacher education programs, the subjects studied (e.g., German, English, mathematics etc.) are classified accordingly. By indicating that a teacher education program is involved, the annotators know that several subjects must be classified from the information provided. When translating English titles, we assess each cleaned subject entry to verify if all lemmatized words are English. For this purpose, their occurrence in the NLTK word corpus is checked. If all words belong to the English language, the processed subject entry is translated.

The data is exported to a separate CSV file and then transferred to a Word document (docx). This document is subsequently translated into German using DeepL's document translation feature. Carrying out an automatic translation in Python without this manual intermediate step is not feasible due to constraints on freely available translation APIs. The German version of English study subject data is added as a new variable to the dataset to be classified and, for efficiency reasons, is only used when assigning a code to the German study subject is not possible. If none of the four dictionaries leads to a successful Destatis code assignment, the process is repeated with the information translated from English to German, if available. Abbreviations present an additional challenge. If students only enter an abbreviation they are familiar with, this can often not be assigned. While common abbreviations such as BWL (for "Betriebswirtschaftslehre" i.e., Business Economics) are already present in the coded data (D4), this is not the case for many other abbreviations. These are problematic when searching for partial matches: If an abbreviation consists of only 2 or 3 letters, the probability is very high that it will also appear in data unrelated to the actual subject. The abbreviation IB ("Internationale Beziehungen", International Relations) would erroneously result in a partial match (as discussed in the following section) with the degree program Library Science. Therefore, abbreviations are manually researched and assigned if they are not contained in the known coded data.

Assigning codes to preprocessed study programs

After both the training dictionaries and the answers from the new survey have been preprocessed, we perform an assignment of the correct numerical codes in a three-step procedure: First direct, fully automated assignments; second, a semi-automated assignment based on automatically generated suggestions; and third, manual assignments for difficult and/or ambiguous cases.

Direct assignment

In the first step, a direct assignment is attempted using the preprocessed version of the subject value by finding values with perfect or almost-perfect matches to known and categorized subjects. A perfect match means the preprocessed text is identical to the reference value from the dictionary, while an almost-perfect match has a Levenshtein distance of one, with equal weights of one each for all delete, insert, and change operations. This is done by comparing the edited version of D1 with the dictionaries created from D2–D4. If this fails and the subject title is in English, a direct assignment is attempted again using the German translation. As soon as one of the variants listed below matches an entry of the dictionaries, the subsequent variants are no longer executed. The dictionaries are searched in the order specified here in advance.

- *Perfect match*: The subject of study is included in one of the four dictionaries.
- *Almost perfect match*: The subject so closely resembles an entry from the four dictionaries that you would only need to swap/remove/insert one letter to match the entry. This case catches many simple typos or slight variations in spelling (e.g. "Gender studies" vs. "Gender Studien" or "matematics" vs. "mathematics").

- *Direct match without spaces*: Since punctuation marks and numbers were replaced by single spaces during the cleanup, it is possible that there is no direct match due to deviating spaces. For this reason, all blanks are removed from both the dictionary entries and the subject entries, and then a direct match is searched for again.
- *Near match without spaces*: Here, all blanks are also removed and an edit distance of 1 is allowed between the entries and the subject specification. In the subsample passed, 270,701 (i.e., 83.6%) of 323,791 non-blank specifications will be directly coded in the first run of the tool.

Uncertain prediction for manual coding

In many cases, however, the identified matches do not allow for a clear assignment to subjects from the list of the Federal Statistical Office. For example, the assignment for “ecotrophology” does not work because the known data only contain the official subject designation “nutrition science”. However, there are some cases where an assignment seems possible with a larger margin of error. Since false-positive assignments are possible here, the variants listed below are used as suggestions and not as direct code assignments. Therefore, all variants are always executed to provide several suggestions if necessary. For example, if “Teaching English and German” is specified as a subject, both English (English Studies) and German (German Studies) should be suggested as subjects. The suggestions resulting from each variant are sorted so that the most frequent suggestion is displayed first. Just as with direct matches, a direct match is first attempted using the prepared German subject entry. If this fails and the subject title is in English, a direct match is attempted again using the German translation. Finally, the suggestions are also exported to a list where they can be checked manually.

- *Split at “and” and direct match*: Some study programs with two subjects are listed by students with an ‘and’ (“und”) connected in only one entry (see previous example). Therefore, it is first tested whether an ‘and’ is included in the entry. If this is the case, an attempt is made to find a direct match for both partial entries to the left and right of the ‘and’.
- *Partial match*: Some specialized study programs (e.g. applied computer science – computer science) or abbreviations (e.g. mathematics – math) allow an assignment via a partial match. To maintain a low false positive rate, a suggestion is made based on whether the student's subject specification is either completely contained in a dictionary entry or vice versa.
- *Almost match*: While in the direct matches only an edit distance of 1 was allowed, i.e., a deviation of only one letter, in the suggestions an edit distance of 2 is allowed. This allows for twisted letters (e.g. chemistry – “Chemie” as “chmeie”). However, since some subjects already sound very similar and the false positive rate would again be too high with a direct assignment (e.g. geology vs. biology), an edit distance of 2 is only used for suggestions. To facilitate the examination of the suggestions, a screen tool was developed and used in which the individual plain text statements with the suggestions generated from the dictionaries are displayed on one screen page each. The codes to be made could be selected and manually coded by clicking the option that seems to apply most probably. The assignments based on the proposals were made successively in the subsample used, i.e. in several rounds. As a result, the remaining manual coding effort could be significantly reduced, i.e. a large part of the Destatis code could be assigned automatically (for details see Table 2).

Non-assignable study program data

If none of the procedures succeeds, the assignment is exported to another sorted list and has to be coded manually without any further clues. If an assignment was made manually, it is logged, just like the suggested assignment, and adds to the total corpus of assignments. In total, only 2,359 plain language entries from the partial data set were coded completely manually.

Merging the results

Applying the classification results in three output files that need to be remerged into the original data set. After manually correcting and coding the suggested and non-assignable study programs in Excel or LimeSurvey (a script for automatically creating a survey and converting its results to the required format is available), the codes need to be inserted into the original data. For this purpose, the classification procedure is run a second time with the updated dictionaries which now include the correct coding for all study program names. We add the name of the identified field of study from the Destatis classification scheme and the corresponding code in new variables to the original data. Since we observed many cases where students listed two fields of study or the two provided fields were not enough, e.g., for some "Lehramt" (teaching degree) students or when having two minors, we allow for a different number of variables than in the original data. These cases are easily identifiable by two or more correct codes from manual or automated coding, although they will only be coded automatically if they were listed with two correct codes in one of the dictionaries. The codes will then be assigned to the new variables sequentially. For clarification purposes, consider the following abstract example (see Table 1) where researchers provided two study program variables in the survey, but decided to create three for the analysis. Student A provided one subject in the first variable and two in the second, student B provided four in the first and none in the second and student C only one in the first and none in the second.

Table 1: Coding examples of original entry fields into Destatis code

ID	original entry 1	original entry 2	field of study 1	field of study 2	field of study 3
A	math	chemistry and biology	math	chemistry	biology
B	politics, communication, history, sociology		politics	communication	history
C	computer science		computer science		

Source: Higher Education Research Group, University of Konstanz

Our procedure assigns the first field of study to the first new variable, the second given field of study to the second new variable and so on. If there are more fields of study than new variables, the overflow will be lost (e.g. sociology from student B in this example). In practice, the codes will of course also be included in the merged data.

How to adapt the procedure to another use case

Our method is easily transferable and adaptable to other data and/or applications. Think of some kind of coded data in the same or a very similar format, e.g., from previous surveys, scraped from the web, synthetically generated, or from any other source. Also, the method is most useful when the data that should be coded has the following characteristics:

- a high number of categories
- it requires expert knowledge to perform the classification
- there is no further context, i.e., only one word or only one noun phrase
- there are a high number of spelling errors and/or other inconsistencies

Our method is published on GitHub and can directly be used to classify study programs. We provide the necessary training data resources (D2–D4) and a small example data set for this case. However, you can easily adapt it to other use cases by making two adjustments: First, modify the underlying resources, i.e., preprocessed training data, commonly used abbreviations, text to remove etc. Second, adjust the specifications in the code for data preprocessing and column selection etc. Detailed instructions are available in the repository. After adapting and running the procedure, you will receive three output files, two of which for manual recoding. You can merge your coded results into a final data set as described before.

Evaluation on final prediction data set

The final data set contains 274,466 observations with 8 field-of-study variables. Note that the coding processes are run for the current and (potentially) former enrollments. Not all respondents provide information on their fields of study. Table 2 shows the distribution of completed fields for all 8 variables. In total, approximately 324,000 statements have been made, the majority of which are in the current major variable. 3,373 statements were additionally identified as missing by preprocessing (statements – statements adjusted). Data (in %) and adjusted data (in %) refer to the percentage of respondents who made a statement for this variable, which is why the sum of the percentage points is greater than 100. The unique data depict how many different values were given by students after all duplicates were eliminated. Unique and adjusted shows the reduction in this number of distinct values due to the preprocessing we performed. The proportion thus maps the proportion at which values from the unique statement list (proportion unique (%)) or the classification procedure input (proportion unique adjusted (%)) occur in the original variable. Since these values can occur in several variables, the sum of the proportions is also greater than 100 or the sum of the specifications is greater than the number of unique specifications specified below. Through preprocessing, we obtain 27,030 unique adjusted values that serve as input for classification, i.e., whose study subject code must be determined.

Table 2: Distribution of completed fields for all 8 variables with field of study entries

variable with open answers	data		adjusted data		unique data		unique and adjusted data	
	abs.	in %	abs.	in %	abs.	in %	abs.	in %
sfach1o2	200,816	73.2	198,969	72.5	31,007	61.2	16,271	60.2
sfach2o2	38,319	14.0	38,091	13.9	10,599	20.9	6,509	24.1
fach01	22,725	8.3	22,137	8.1	7,485	14.8	5,043	18.7
fach02	19,394	7.1	18,940	6.9	7,236	14.3	5,068	18.7
fach03	7,137	2.6	7,015	2.6	3,364	6.6	2,622	9.7
fach04	1,666	0.6	1,631	0.6	1,023	2.0	855	3.2
sabserfacho1	27,169	9.9	27,096	9.9	6,189	12.2	4,321	16.0
sabserfacho2	6,565	2.4	6,539	2.4	2,019	4.0	1,527	5.6
Total	323,791		320,418		68,922		42,216	

Source: Higher Education Research Group University of Konstanz;
Note: *sfach1o2*: first field of study; *sfach2o2*: [potentially] second field of study; *fach01*: field of study of first episode of enrollment history; *fach02*: field of study of second episode of enrollment history; *fach03*: field of study of third episode of enrollment history; *fach04*: field of study of forth episode of enrollment history; *sabserfacho1*: first field of study of a previously attained degree; *sabserfacho2*: second field of study of a previously attained degree

Table 3 shows the number of specifications. Number of entries indicates how many values in the original data set fall within the classification variant. The next column provides the relative distribution of the entries (in %) in each classification variant. The last two columns contain the absolute and relative numbers of cases preprocessing, removal of duplicates and (most importantly) automated classification. The number of entries that needs to be manually checked significantly reduces to 2,359 cases. Noteworthy, that quite a high number of entries (13,435) is still to be checked on suggestion – in order to reduce false (automatic) classifications.

Table 3: Number of specifications

coding scheme	number of entries	in %	number unique and adjusted	in %
automatically	270,701	83.6	11,236	41.6
on suggestion	38,978	12.0	13,435	49.7
manual	14,112	4.4	2,359	8.7
total	323,791	100	27,030	100

Source: Higher Education Research Group, University of Konstanz

Conclusion

We propose a semi-automated coding procedure to be used to code open-ended natural language on the fields of study to the “structured” Destatis classification scheme (“Systematik der Studienfächer”). After preprocessing of the survey data for about 95 percent of cases, a code was assigned. About 4 percent of cases could be manually coded based on suggestions provided by the procedure. Less than one percent needed manual coding without any suggestions. Since the dictionary developed with correctly assigned codes grows with each application, the share of not-automatically coded data entries will further shrink. The procedure is based on strict matching criteria, the error rate should thus be neglectable. Of course, the dictionaries need to be updated on a regular basis because the assignments in the Destatis code change occasionally, or new study programs are developed.

References

Beuße, M., Kroher, M., Becker, K., Ehrhardt, M.-K., Isleib, S., Koopmann, J., Steinkühler, J., Völk, D., Buchholz, Meyer, J., Multrus, F., Hinz, T., Marczuk, A., & Strauß, S. (2022). Die Studierendenbefragung in Deutschland: Eine neue, integrierte Datenbasis für Forschung, Bildungs-und Hochschulpolitik. (DZHW Brief 06|2022). Hannover: DZHW. https://doi.org/10.34878/2022.06.dzhw_brief

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3, 993–1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Card, D., & Smith, N. A. (2015). Automated Coding of Open-Ended Survey Responses. Retrieved October 20 from https://www.ml.cmu.edu/research/dap-papers/DAP_Card.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [Cs]. https://arxiv.org/abs/1810.04805

Eshima, S., Imai, K., & Sasaki, T. (2023). Keyword-Assisted Topic Models. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12779

Harandizadeh, B., Priniski, J. H., & Morstatter, F. (2022). Keyword Assisted Embedded Topic Model. WSDM '22: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 372–380. https://doi.org/10.1145/3488560.3498518

Krippendorff, K. (1978). Reliability of binary attribute data. *Biometrics*, 34, 142–144.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [Cs]. https://arxiv.org/abs/1907.11692

Lörz, M., Marczuk, A., Zimmer, L., Multrus, F., & Buchholz, S. (2020). Studieren unter Corona-Bedingungen: Studierende bewerten das erste Digitalsemester. (DZHW Brief 5|2020). Hannover: DZHW. https://doi.org/10.34878/2020.05.dzhw_brief

Padurariu, C., & Breaban, M. E. (2019). Dealing with Data Imbalance in Text Classification. *Procedia Computer Science*, 159, 736–745. https://doi.org/10.1016/j.procs.2019.09.229

Pietsch, A.-S., & Lessmann, S. (2018). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, 1(2), 93–116, https://doi.org/10.1080/2573234X.2019.1590131

Schonlau, M., Gweon, H., & Wenemark, M. (2021). Automatic Classification of Open-Ended Questions: Check-All-That-Apply Questions. *Social Science Computer Review*, 39(4), 562–572. https://doi.org/10.1177/0894439319869210

Schröder, C., & Niekler, A. (2020). A Survey of Active Learning for Text Classification using Deep Neural Networks. ArXiv, abs/2008.07267.

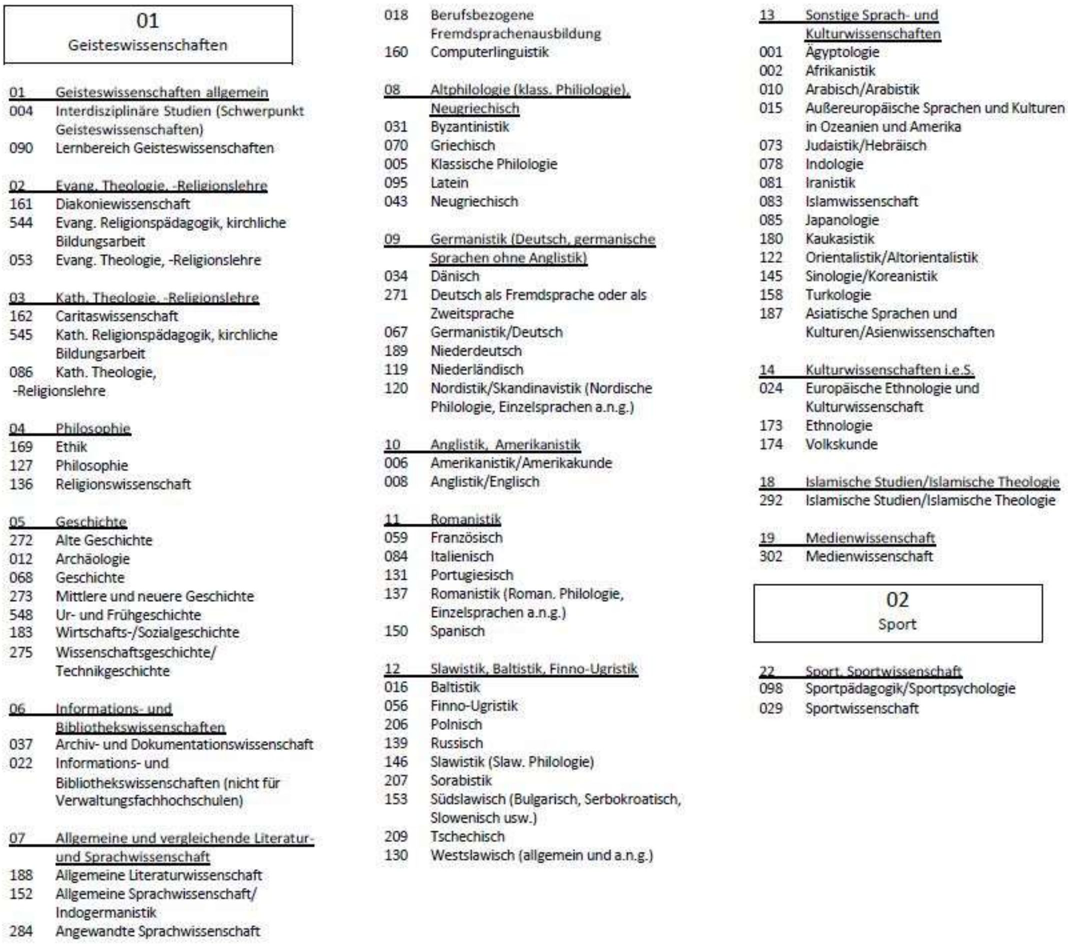
Statistisches Bundesamt (2023). Bildung und Kultur. Studierende an Hochschulen – Fächersystematik 2021. Retrieved October 20, 2023 from https://www.destatis.de/DE/Methoden/Klassifikationen/Bildung/studenten-pruefungsstatistik.pdf?__blob=publicationFile

StudyCheck (n.d.). Retrieved October 20, 2023, from https://www.studycheck.de/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Appendix

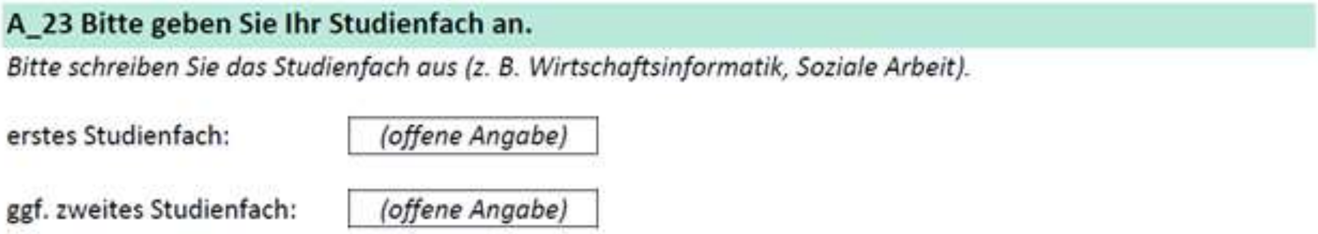
Figure 1A: German version of the Destatis classification scheme (Statistisches Bundesamt, example: humanities)



01	Geisteswissenschaften
01	Geisteswissenschaften allgemein
004	Interdisziplinäre Studien (Schwerpunkt Geisteswissenschaften)
090	Lernbereich Geisteswissenschaften
02	Evang. Theologie, -Religionslehre
161	Diakoniewissenschaft
544	Evang. Religionspädagogik, kirchliche Bildungsarbeit
053	Evang. Theologie, -Religionslehre
03	Kath. Theologie, -Religionslehre
162	Caritaswissenschaft
545	Kath. Religionspädagogik, kirchliche Bildungsarbeit
086	Kath. Theologie, -Religionslehre
04	Philosophie
169	Ethik
127	Philosophie
136	Religionswissenschaft
05	Geschichte
272	Alte Geschichte
012	Archäologie
068	Geschichte
273	Mittlere und neuere Geschichte
548	Ur- und Frühgeschichte
183	Wirtschafts-/Sozialgeschichte
275	Wissenschaftsgeschichte/ Technikgeschichte
06	Informations- und Bibliothekswissenschaften
037	Archiv- und Dokumentationswissenschaft
022	Informations- und Bibliothekswissenschaften (nicht für Verwaltungsfachhochschulen)
07	Allgemeine und vergleichende Literatur- und Sprachwissenschaft
188	Allgemeine Literaturwissenschaft
152	Allgemeine Sprachwissenschaft/ Indogermanistik
284	Angewandte Sprachwissenschaft
018	Berufsbezogene Fremdsprachenausbildung
160	Computerlinguistik
08	Altphilologie (klass. Philiologie), Neugriechisch
031	Byzantinistik
070	Griechisch
005	Klassische Philologie
095	Latein
043	Neugriechisch
09	Germanistik (Deutsch, germanische Sprachen ohne Anglistik)
034	Dänisch
271	Deutsch als Fremdsprache oder als Zweitsprache
067	Germanistik/Deutsch
189	Niederdeutsch
119	Niederländisch
120	Nordistik/Skandinavistik (Nordische Philologie, Einzelsprachen a.n.g.)
10	Anglistik, Amerikanistik
006	Amerikanistik/Amerikakunde
008	Anglistik/Englisch
11	Romanistik
059	Französisch
084	Italienisch
131	Portugiesisch
137	Romanistik (Roman. Philologie, Einzelsprachen a.n.g.)
150	Spanisch
12	Slawistik, Baltistik, Finno-Ugristik
016	Baltistik
056	Finno-Ugristik
206	Polnisch
139	Russisch
146	Slawistik (Slaw. Philologie)
207	Sorabistik
153	Südslawisch (Bulgarisch, Serbokroatisch, Slowenisch usw.)
209	Tschechisch
130	Westslawisch (allgemein und a.n.g.)
13	Sonstige Sprach- und Kulturwissenschaften
001	Ägyptologie
002	Afrikanistik
010	Arabisch/Arabistik
015	Außereuropäische Sprachen und Kulturen in Ozeanien und Amerika
073	Judaistik/Hebräisch
078	Indologie
081	Iranistik
083	Islamwissenschaft
085	Japanologie
180	Kaukasistik
122	Orientalistik/Altorientalistik
145	Sinologie/Koreanistik
158	Turkologie
187	Asiatische Sprachen und Kulturen/Asienwissenschaften
14	Kulturwissenschaften i.e.S.
024	Europäische Ethnologie und Kulturwissenschaft
173	Ethnologie
174	Volkskunde
18	Islamische Studien/Islamische Theologie
292	Islamische Studien/Islamische Theologie
19	Medienwissenschaft
302	Medienwissenschaft
02	Sport
22	Sport, Sportwissenschaft
098	Sportpädagogik/Sportpsychologie
029	Sportwissenschaft

Source: Statistisches Bundesamt 2023, German version „Studierende an Hochschulen – Fächersystematik“

Figure 2A: Open-ended question on the field of study (“Studienfach”), screenshot German questionnaire



A_23 Bitte geben Sie Ihr Studienfach an.

Bitte schreiben Sie das Studienfach aus (z. B. Wirtschaftsinformatik, Soziale Arbeit).

erstes Studienfach: *(offene Angabe)*

ggf. zweites Studienfach: *(offene Angabe)*

Source: SiD, Beuße et al. 2022

Impressum

Herausgeben von
Universität Konstanz
AG Hochschulforschung
Universitätsstraße 10
78464 Konstanz
vertreten durch Rektorin Prof. Dr. Katharina Holzinger
uni-konstanz.de

Stand
Februar 2024

Text, Grafiken und Gestaltung
AG Hochschulforschung, Universität Konstanz

ISSN 1616-0398