Datenbeschreibung

RWI – Leibniz-Institut für Wirtschaftsforschung

FDZ Data Description: RWI-UNI-SUBJECTS: Complete Records of All Subjects Across German HEIs (1971–1996)

Friederike Hertweck, Lukas Jonas, Boris Thome and Serife Yasar

Revised edition – July 2025
(replaces version from November 2024)

Impressum

Herausgeber:

RWI – Leibniz-Institut für Wirtschaftsforschung

Hohenzollernstraße 1–3 | 45128 Essen, Germany

Postanschrift

Postfach 10 30 54 | 45030 Essen, Germany

Fon: +49 201–81 49-0 | E-Mail: rwi@rwi-essen.de
www.rwi-essen.de

Vorstand

Prof. Dr. Dr. h.c. Christoph M. Schmidt (Präsident)

Prof. Dr. Thomas K. Bauer (Vizepräsident)

Dr. Stefan Rumpf (Administrativer Vorstand)

Prof. Dr. Kerstin Schneider (Mitglied des erweiterten Vorstands)

RWI Datenbeschreibung

Schriftleitung: Prof. Dr. Dr. h. c. Christoph M. Schmidt

Gestaltung: Magdalena Franke, Claudia Lohkamp

FDZ Data Description: RWI-UNI-SUBJECTS: Complete Records of All Subjects Across German HEIs (1971–1996)

Revised edition – July 2025
(replaces version from November 2024)

Friederike Hertweck, Lukas Jonas, Boris Thome and Serife Yasar

Leibniz-Institut für Wirtschaftsforschung

Leibniz Gemeinschaft

Das RWI wird vom Bund und vom Land Nordrhein-Westfalen gefördert.

RWI-UNI-SUBJECTS:
Complete records of all subjects across German HEIs (1971-1996)

Friederike Hertweck[*†] Lukas Jonas[‡] Boris Thome[§] Serife Yasar[¶]

July 22, 2025

Abstract

The dataset RWI-UNI-SUBJECTS provides comprehensive information on the universe of academic subjects offered by the universe of higher education institutions across Germany between 1971 and 1996. RWI-UNI-SUBJECTS is based on tables extracted from paper-based guides on "Study and Career Choice" that are annually published by the Federal Employment Agency to inform high school students about post-secondary education. The dataset is available on request as a scientific-use file from the Research Data Center Ruhr (FDZ Ruhr). It is structured to easily link locations, universities, and subjects with administrative and survey data. This dataset holds significant potential for use across various disciplines, including history, sociology, political science, economics, and other fields analyzing the higher education market.

Keywords: Higher Education, Colleges, Educational Expansion, Dataset, Germany

JEL Codes: I20, I23, J24

[*]RWI – Leibniz Institute for Economic Research, Essen, Germany

[†]The authors thank Fabian Dell, Christoph Dickmann, Alina Niemann, Marc Schmidt, and Yang Shen for their support in compiling the dataset. The creation of the RWI-UNI-SUBJECTS dataset is part of the project "NLP in Education Research: Development of the German higher education market since the 1970s (Studi-BUCH)" funded by the Federal Ministry of Research, Technology and Space (BMFTR) and Next Generation EU under grant numbers 16DKWN139A and 16DKWN139B.

Correspondence: friederike.hertweck@rwi-essen.de.

[‡]RWI – Leibniz Institute for Economic Research, Essen, Germany & Ruhr-University Bochum, Bochum, Germany

[§]Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[¶]RWI – Leibniz Institute for Economic Research, Essen, Germany & Ruhr-University Bochum, Bochum, Germany

Contents

1 Overview and Analytic Options

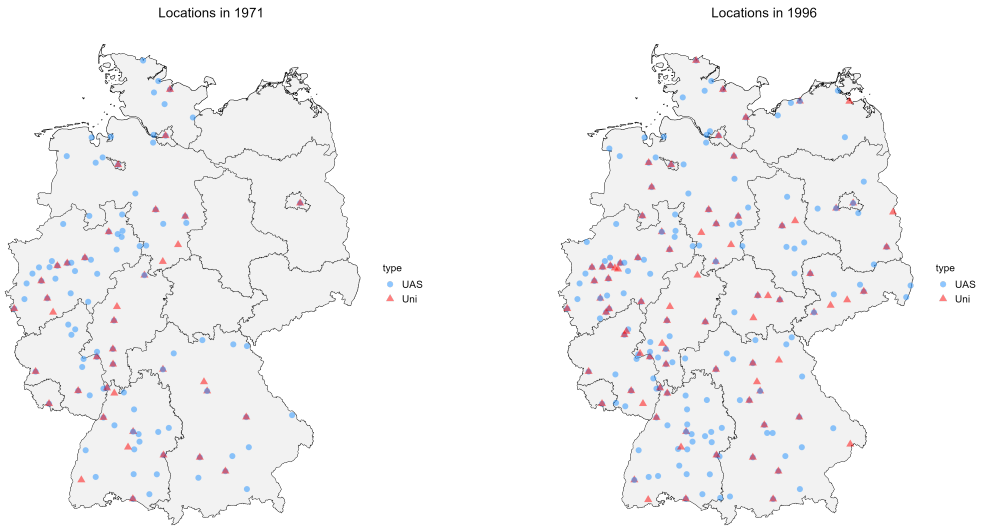
1.1 Introduction

The dataset RWI-UNI-SUBJECTS is based on the guide on "Study and Career Choice" (in German: *"Studien- und Berufswahl"* and recently *"Studienwahl"*). The guides provide high school graduates and prospective students with comprehensive information on choosing a higher education program in Germany. The book is considered as the official guide to studying in Germany and is provided and distributed on behalf of the Federal Employment Agency. The book contains detailed information on higher education in Germany, including information on institutions, subjects, study opportunities, financial aid, student housing, and related topics.

The first guide was published in 1971 and has been updated yearly. Based on digitized copies of the guides covering the years 1971 to 1996,[1] the dataset documents all universities and universities of applied sciences across Germany during this period, along with the array of subjects available for study at these institutions on a yearly basis.

Figure 1 illustrates the geographic distribution of higher education institutions for the years 1971 and 1996. It is worth noting that data on the higher education landscape of the German Democratic Republic (GDR) is not available. As a result, the eastern states of Germany were integrated into the dataset in 1991, the year subsequent to reunification. Any information regarding locations in Eastern Germany before 1991 is thus absent from the dataset. Moreover, alongside the establishment of higher education institutions in the newly formed federal states following reunification, Figure 1 also highlights a notable surge in the number of higher education institutions within the old federal states, particularly evident in regions such as Lower Saxony and Bavaria.

Figure 1: Higher education locations in Germany in 1971 and 1996



Note: Data on the higher education landscape of the German Democratic Republic is not available. Thus, data pertaining to institutions in the Eastern states is only available from the year 1991 onwards.

Depending on the year, it is possible to distinguish 89 to 169 different subjects (in German: *"Studienfach"*). These are grouped into up to 61 different subject areas (in German: *"Studienbereiche"*) and eight broad subject groups (in German: *"Fächergruppe"*). These subject groups are: Humanities (01),

[1] From 1971 to 1996, the study guides followed a standardized presentation of the study programs. Afterwards, the design changed so that it is not possible to extract the same information from the books from 1997 onwards with the methodology used so far.

Sports (02), Law, Economics and Social Sciences (03), Mathematics and Natural Sciences (04), Human Medicine and Health Sciences (05), Agricultural, Forestry, and Nutritional Sciences, as well as Veterinary Medicine (07), Engineering (08), and Arts and Art Sciences (09). The aggregation of subjects follows the academic coding system provided by Destatis (2023). Please also refer to Appendix A for a more detailed description of the fields.

1.2 Analytic Options

The RWI-UNI-SUBJECTS exhibits significant potential for addressing various research questions due to its comprehensive coverage of the expansion of higher education in Germany over 25 years. The dataset provides insights into the evolution of the German higher education landscape and tracks fluctuations in the number of higher education institutions across both time and regions, capturing the dynamic progression of academic subjects. Thus, it enables nuanced analyses of the expansion of specific subjects, subject areas, and subject groups, thereby facilitating examinations of institutional dynamics, subject trends, and regional shifts starting from the 1970s onwards.

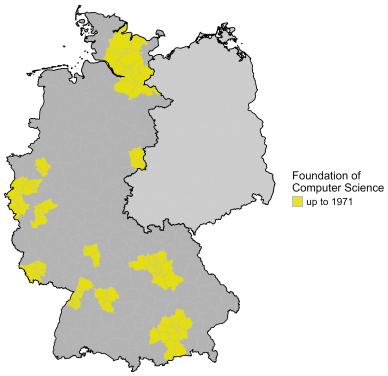
Furthermore, the dataset offers opportunities for linkages with other datasets through its temporal and regional attributes. In order to facilitate regional analyses, the official municipality code (in German: "Amtlicher Gemeindeschlüssel, AGS") has been integrated into the dataset. The municipality code enables seamless connection with educational panel datasets, such as the National Educational Panel Study (NEPS-Netzwerk, 2023), the Sample of Integrated Employment Biographies (SIAB) provided by the Federal Employment Agency (IAB) (Schmucker, Seth, & vom Berge, 2023), or the DZHW Graduate Panel 1989 (German Centre for Higher Education Research and Science Studies (DZHW), 2020). These linkages enhance the dataset's utility and extend its potential for comprehensive research endeavors across various domains.

To the best of our knowledge, only two datasets exist so far that are slightly comparable to RWI-UNI-SUBJECTS but are less exhaustive. First, the *College Scorecard dataset* (U.S. Department of Education, 2024) contains comprehensive institution-level data from 1996 to 2023 on enrollment, financial aid, costs, debt, repayment, and post-graduation earnings as well as the same information on the subject-level for 2014 to 2020. It also includes crosswalk files that link colleges' identification codes (Office of Postsecondary Education Identification/OPEID) with identifiers from the Integrated Postsecondary Education Data System (IPEDS UnitID).

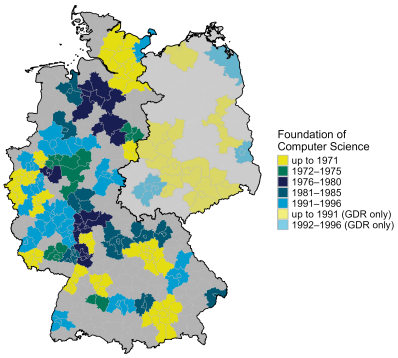
Second, the *Catalogue of First and Second Cycle Degree Programmes of the University of Bologna* (University of Bologna, 2024) provides detailed program data by academic year, covering the period from 2004/2005 to 2024/2025, exclusively for the University of Bologna. Yet, neither dataset fully covers or substitutes for the RWI-UNI-SUBJECTS dataset, as both provide information on only a subsample of institutions, cover different and in particular much shorter time periods, and thereby do not address the years of higher education expansion in the Western world throughout the second half of the 20^{th} century.

Overall, the RWI-UNI-SUBJECTS dataset offers unique insights into the development of German higher education from 1971 to 1996. The period is marked by substantial expansion and educational and societal transformation, both within Germany and across Western higher education more broadly. Although comparable data is currently not available beyond 1996, the RWI-UNI-SUBJECTS dataset remains a valuable resource for understanding the historical distribution of study fields across institutions and regions. Its longitudinal perspective continues to inform current analyses of labor markets, regional development, and patterns of industry specialization. Figure 2 specifically illustrates evolution of Computer Science at German higher education institutions across labor market regions during this period.

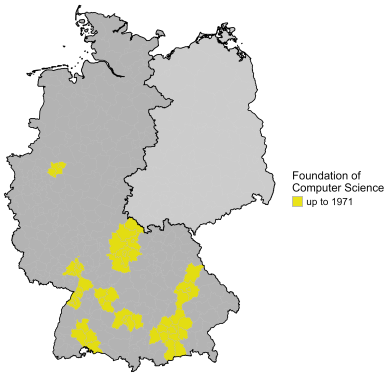
Figure 2: Evolution of computer science (CS) at German higher education institutions



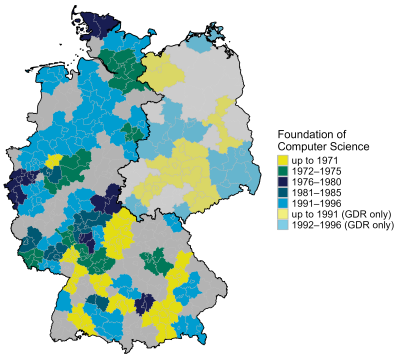
A) CS at universities in 1971



B) CS at universities until 1996



C) CS at UAS in 1971



D) CS at UAS until 1996

1.3 Descriptives

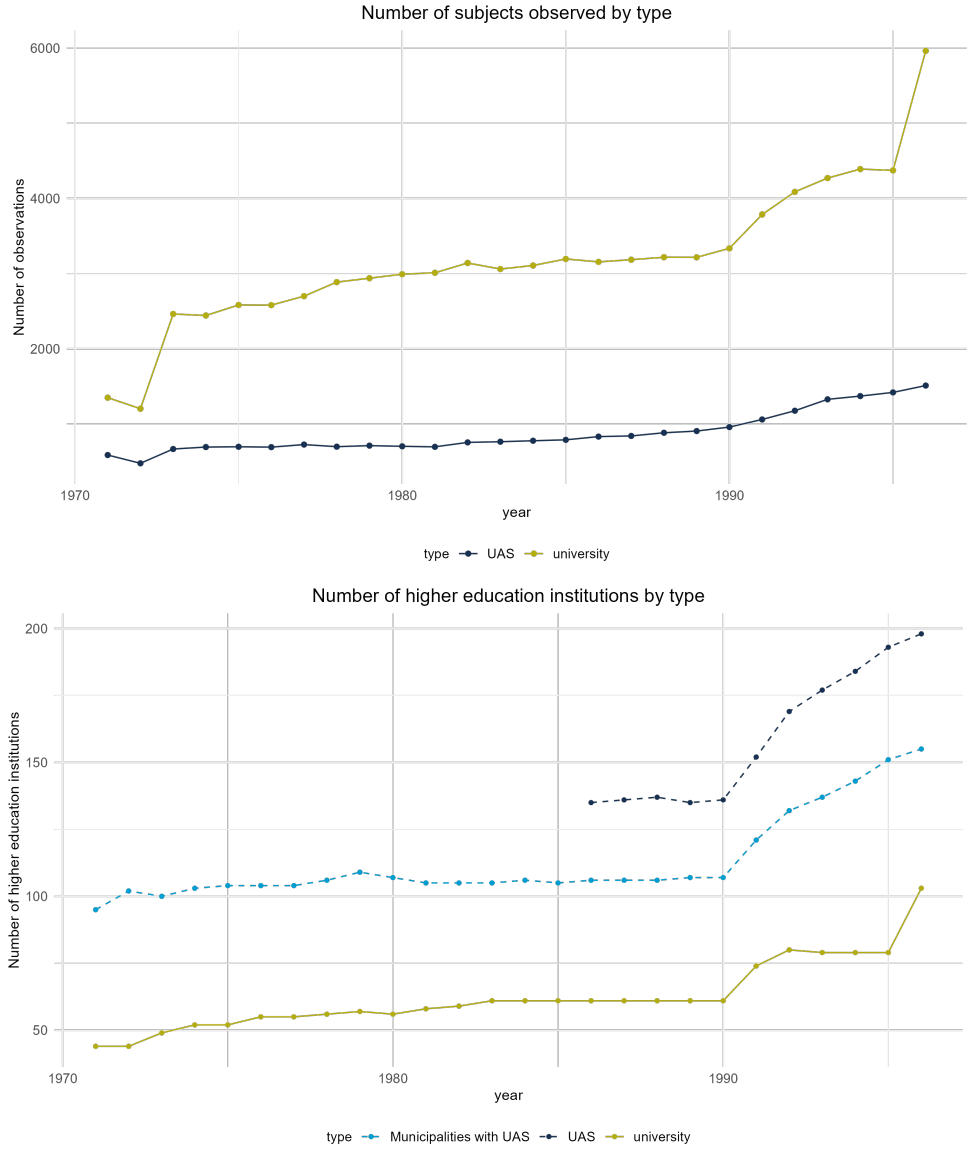
Between 1971 and 1996, both the number of higher education institutions and the number of subjects increased. The upper panel of Figure 3 illustrates the temporal evolution of the total number of subjects available at universities and UAS. In 1971, students had a choice among 1,349 subjects at universities and 586 subjects at UAS. By 1990, these figures had risen to 3,335 subjects at universities and 958 subjects at UAS. By 1996, the total number of subjects had further increased to 5,962 at universities (including locations in the former GDR) and 1,510 subjects at UAS.

The lower panel of Figure 3 illustrates the temporal evolution of the total number of higher education institutions. In 1971, students could enroll in one of 44 universities, a figure that had increased to already 61 by 1990 and reflects the expansion of universities in Western Germany. By 1996, the number of universities in the data had further risen to 103 and includes also those universities of the former GDR.

Any information on UAS was aggregated at the city level in the guides until 1985, after which it was listed separately for each UAS. By 1996, the number of UAS increased to 198 from 135 in 1986. Notably, even excluding the 1991 spike attributable to the integration of the former GDR, the data show a strong expansion in both institutions and subjects over the period covered in RWI-UNI-SUBJECTS.

Overall, the dataset encompasses all study options at universities and universities of applied sciences documented in the guides on "Study and Career Choice". It is, however, possible that the guides' editors inadvertently omitted certain study options. Thus, any mistakes in the official guides may have been brought forward into the dataset and are discussed in section 2.4.

Figure 3: Temporal evolution of higher education landscape



Note: The upper panel presents the number of subjects separately for UAS and universities. The lower panel provides the number of higher education institutions over time. In the lower panel, universities are represented in yellow, UAS in dark blue, and aggregated UAS information at the city level is depicted in light blue until 1985. Prior to 1986, if two or more UAS were situated in the same city, the guides aggregated the information at the city level without distinguishing between individual UAS. However, a clear differentiation between separate UAS is provided from 1986 onward. Underlying numbers are available in Tables 1 and 2 in Appendix B.1.

1.4 Data access

The dataset is available as a scientific-use file via the Research Data Center Ruhr (FDZ Ruhr). Forwarding the data without the consent of the FDZ Ruhr is prohibited. The FDZ Ruhr expects the dataset to be used responsibly. The dataset must be cited as follows:
Hertweck, F., Jonas, L., Thome, B., and Yasar, S., 2024. RWI-UNI-SUBJECTS: Complete records of all subjects across German HEIs (1971 - 1996). RWI-Micro. Version: 1. RWI – Leibniz Institute for Economic Research. Dataset. doi:10.7807/studi:buch:suf:v1.

2 Data preparation

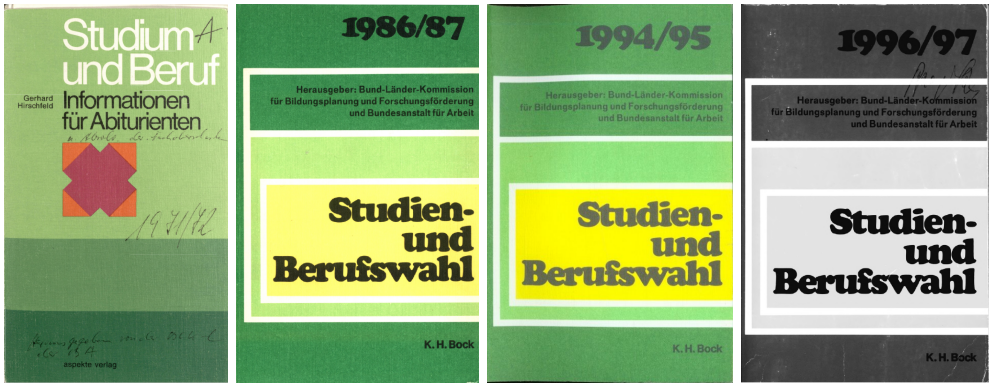
The creation of the RWI-UNI-SUBJECTS dataset involves several key steps. It begins with converting PDF tables into images, followed by the use of OpenCV to recognize table grids and symbols. To ensure the integrity of the extracted information, extensive accuracy checks are conducted, including visual reviews and comparisons with original sources. Furthermore, additional data preparation steps involve correcting the names of higher education institutions and subjects to align them with unique identifiers from the German Federal Statistical Office, and incorporating official municipality codes to facilitate seamless connection with other datasets.

2.1 Data source

The dataset RWI-UNI-SUBJECTS is based on tables from the guides on "Study and Career Choice" (in German: *"Studien- und Berufswahl"*) from 1971 to 1996, provided by the Federal Employment Agency (Hirschfeld, 1971, 1972; Bock, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996). These guides have been digitised and stored in *Portable Document Format* (PDF for short).

These guides inform high school students towards the end of upper secondary school about higher education and occupational career paths in Germany. Each year, high schools distribute these guides to their students free of charge. At the time of publication, the Federal Employment Agency (in German: *"Bundesagentur für Arbeit"*) describes these guides as the "official study guide for Germany" (in German: *"offiziellen Studienführer für Deutschland"*). Due to its color, it is also known as "the green book" (see Figure 4).

Figure 4: Cover pages of the guides from 1971, 1986, 1994 and 1996



Note: A sample of cover pages. These are based on Hirschfeld (1971) and Bock (1986, 1994, 1996).

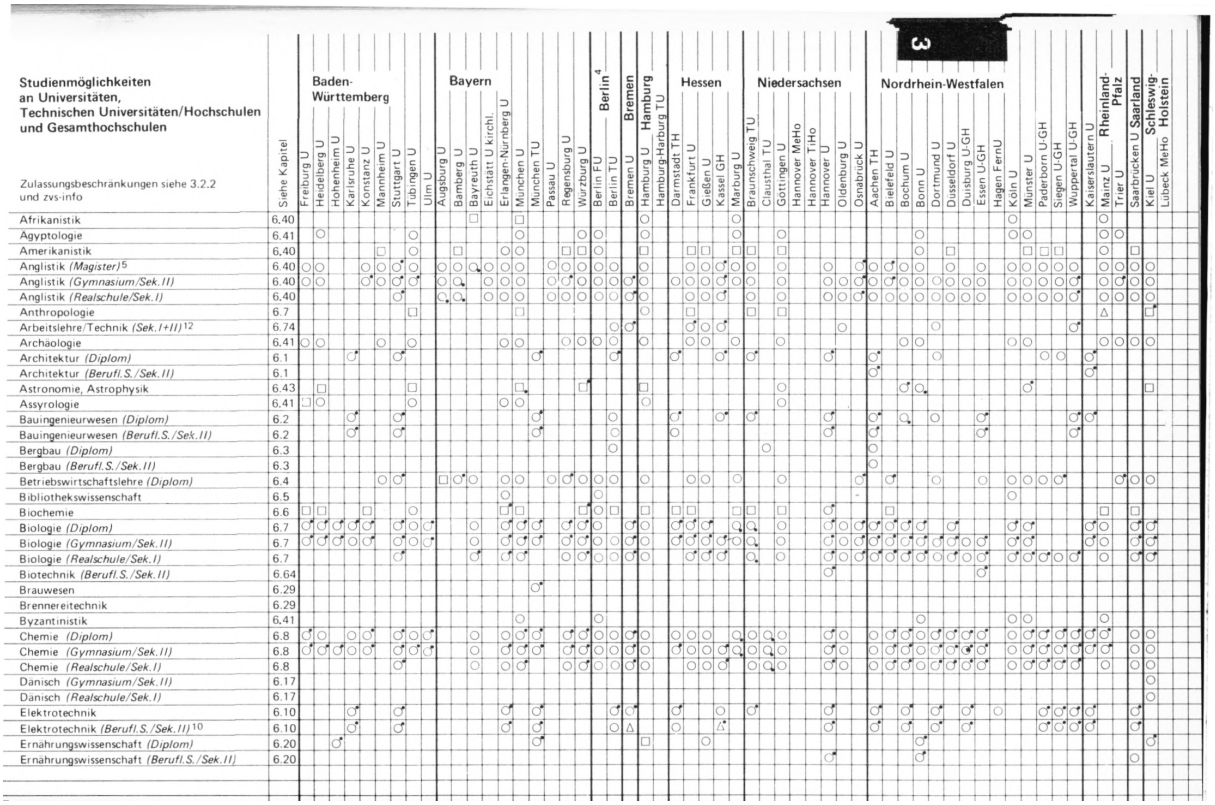
The guides have been distributed to high school students from 1971 onwards. The state monopoly mandate for guidance on career choice (in German: *"staatlicher Monopolauftrag einer öffentlichen Berufsberatung"*) was confirmed in the Labor Promotion Law (in German: *"Arbeitsförderungsgesetz"*) on 25 June 1969 and transferred to the Federal Labor Agency (formerly known as *"Bundesanstalt für Arbeit"*, now *"Bundesagentur für Arbeit"*). This law created the institutional framework for uniform organization and development of career guidance in Germany (Meisel, 1978). Since the first edition in 1971, the guides to "Study and Career Choice" are annually updated, published, and distributed to high school students. Over the past 50 years, several publishers have been involved in the publishing process.

Each of these guides provides comprehensive details on all subjects available across Germany and the higher education institutions offering them. Specifically, these guides comprise tables that summarize the

availability of subjects across various locations (see Figure 5). In some instances, even second campuses of higher education institutions are listed separately within these tables.

These tables consistently adhere to a standardized structure across all books published between 1971 and 1996. Columns correspond to locations, universities, or UAS (from books starting in 1986 onwards), while rows represent different subjects. The second column denotes the chapter in the book where additional information related to each topic can be further accessed. Information regarding the availability of a subject, program type (e.g., full-time or part-time), the starting term (e.g., winter or summer term), and entry requirements are succinctly summarized using different symbols within the tables. Figure 5 illustrates an extract from one of these tables from the year 1980 (Bock, 1980). Each table spans multiple pages.

Figure 5: Example page of a subject table at universities (1980)



Note: Extract of the table of subjects at universities of the year 1980, based on Bock (1980). The structure of these tables has remained standardized throughout the years.

2.2 Automated extraction of tables

Pre-processing of tables

In the first step, the relevant pages of the PDF files were converted into image files. These image files were saved as *Portable Network Graphics* (PNG for short) with a pixel density of 300. Subsequently, Python-based image processing functions from *OpenCV* were utilized for further processing (Bradski, 2000).

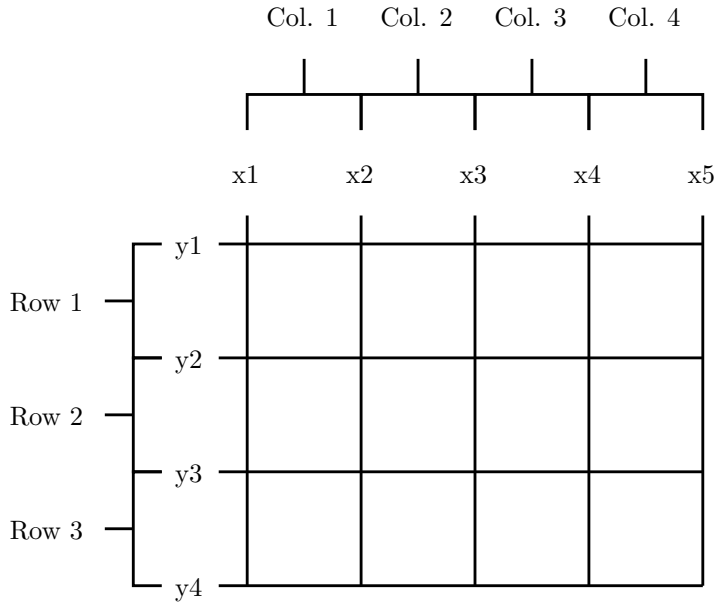
The interior area of the tables within the image files was isolated and, if necessary, rotated to ensure that only the symbols and the table grid (without axis labels) were visible. This approach offers the advantage of preventing axis labels from being erroneously interpreted as table contents. The axis labels

were separately recorded for each table and subsequently integrated into the grids through automated procedures.

Recognition of table grid and symbols

The extraction of the tables can be divided into two tasks: recognizing the table grid and extracting the symbols within the cells. To recognize the **table grid**, OpenCV functions were employed, leveraging horizontal and vertical line recognition facilitated by a kernel. To circumvent thicker intermediate lines being identified as multiple lines, a minimum distance between lines was specified, a value that could vary depending on the respective guide. The identified lines were then arranged in ascending order based on their pixel coordinates and subsequently filtered using the minimum distance. This ensured that the area between two lines corresponded to a row or column. The prior rotation of the tables (if necessary) ensured that horizontal lines commence and conclude at the same y coordinates. Similarly, vertical lines exhibit consistent x coordinates at their top and bottom, eliminating shifts due to rotation. This streamlines the storage of recognized line coordinates and simplifies the subsequent insertion of symbol positions into the table.

Figure 6: Illustration of the table structure



Note: The figure illustrates a schematic example of the structure of the tables. The higher education institutions are listed in columns, while subjects are listed in rows.

Figure 6 illustrates how the coordinates are stored efficiently as a tuple. With this data structure, each row tuple uses two y coordinates and each column tuple uses two x coordinates. For example, the tuple (x_1, x_2) would describe a column that starts at position x_1 and ends at position x_2. The subsequent column (x_2, x_3) starts at the position x_2, i.e., where the preceding column concluded, and terminates at position x_3. The storage of row coordinates follows a similar structure, with y coordinates representing the height position. These tuples, describing the rows or columns, construct a representation of the table grid and are stored in the first row or column of a data frame.

The second step in the process of extracting the information from the tables involves recognizing the various **symbols** for which OpenCV's *template matching*[2] implementation was used.

[2] `https://docs.opencv.org/4.x/d4/dc6/tutorial_py_template_matching.html`

Template matching entails identifying a small pattern or image within a larger image. To implement this procedure, sample images, known as *templates*, are necessary. The template matching algorithm compares a portion of the larger image with the template by systematically moving the template across the entire image. At each position, a comparison is conducted between the template and the corresponding section of the larger image, yielding a similarity value based on color values. If this similarity value surpasses a predefined *threshold*, the template is recognized at the corresponding position. Thus, for the template matching algorithm to operate effectively, at least one sample image is required for each symbol intended for recognition within the table. This sample image should ideally serve as the most accurate representation (e.g., same size, rotation, etc.) of the respective symbol.

Figure 7: Sample of symbol templates

Note: The figure presents several symbols that appear in the tables. The shape of the symbol and the presence and position of the black box within each symbol signify specific attributes such as program type and starting semester.

The templates for the symbols were extracted from the original images (refer to Figure 7) and passed to the template matching algorithm. Given the diversity of symbols in the guides and variations in scan quality, templates were generated separately for university and UAS tables within each guide. In certain instances, employing multiple templates for symbols that posed recognition challenges proved advantageous. Due to the inherent similarity among some symbols, the algorithm selects the symbol with the highest similarity value.

The template matching algorithm returns the x and y coordinates of each recognized symbol along with the similarity value of the evaluated template. If the similarity value exceeds the specified threshold, an associated symbol code and similarity value are stored in the data frame. Data is organized using pre-determined row and column tuples, along with the coordinates of recognized symbols: if symbol coordinates (x, y) fall between column values (x_i, x_{i+1}) and row values (y_j, y_{j+1}), the symbol is placed in column i and row j. This process iterates for each symbol, ensuring only the symbol with the highest similarity value is retained in the data frame.

A complication arose from smaller additional symbols, depicted as small dots adjacent to the main symbol (see Figure 7), that convey secondary meanings, such as *studies begin in the winter term* or *recommended to start in winter term*. Thus, a template icon for each symbol combination was supplied to the algorithm. However, these supplementary symbols, which accompanied a main symbol, were not consistently and precisely identified due to the variability in the spacing between the main symbol and its accompanying dot. Yet, even if the smaller symbol was not recognized, the recognition of the larger symbol remained correct.

In the final step, the algorithm compares the similarity scores for each icon in each cell. Only the numerical code with the highest similarity score is kept in the cells. If no icon template is recognized for a cell, the cell remains empty. This results in a table that holds up to one numerical code in each cell. Axis labels containing subject and location names are then reintegrated into the tables, replacing coordinate tuples for rows and columns. Given that overview tables span multiple pages, tables extracted from individual pages are concatenated into a single comprehensive table for universities and UAS, respectively.

Evaluation of the extraction process

The image-based extraction of the tables from the guides on "Study and Career Choice" proved to be a major challenge. The unique format of symbol-based tables necessitated a tailored technical solution.

Given the dynamic presentation of tables in the guides, the algorithm had to be highly adaptable. Moreover, the quality of the PDF scans varied considerably, with common defects including blurring, rotation, overexposure, or contamination in the scan.

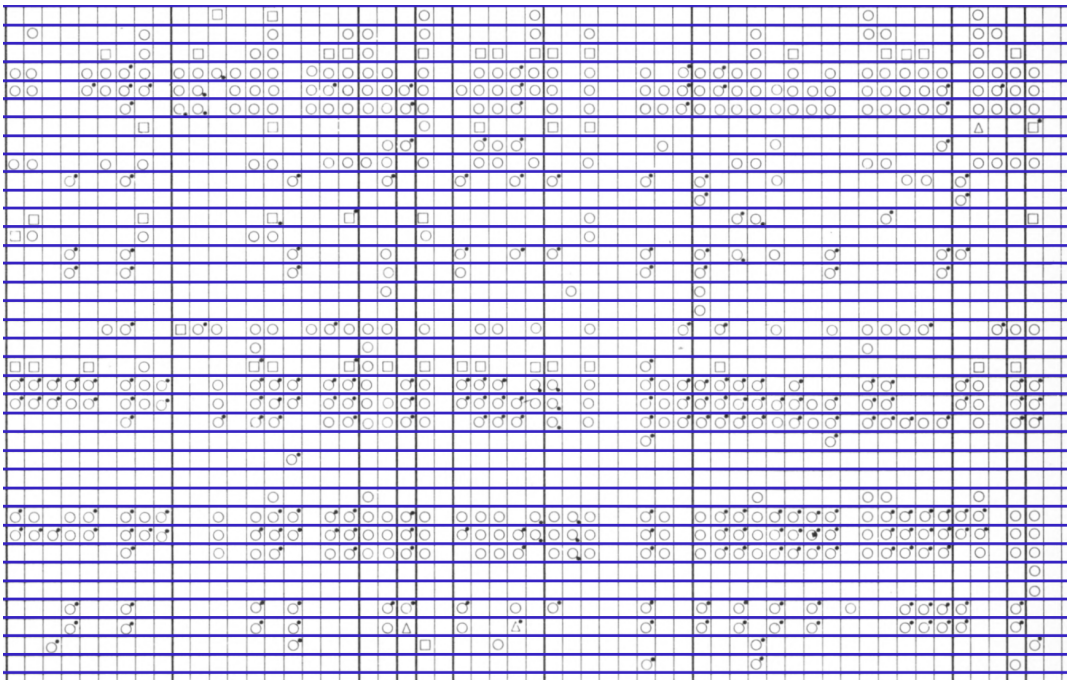
To ensure data quality, the outcomes of the table extraction were thoroughly examined throughout the extraction process. Multiple interim checks were conducted as follows in iterative processes:

- Visual Inspection of intermediate results: Visualizations of the intermediate results were generated, aiding in the identification of any inconsistencies
- Manual comparison between extractions and original: Extracted tables were compared against the original images to verify accuracy and completeness.
- Error Identification: Any errors or discrepancies were promptly identified and flagged for further investigation.

By incorporating these steps into the extraction process, potential issues were promptly addressed, thereby enhancing the overall reliability and accuracy of the extracted data.

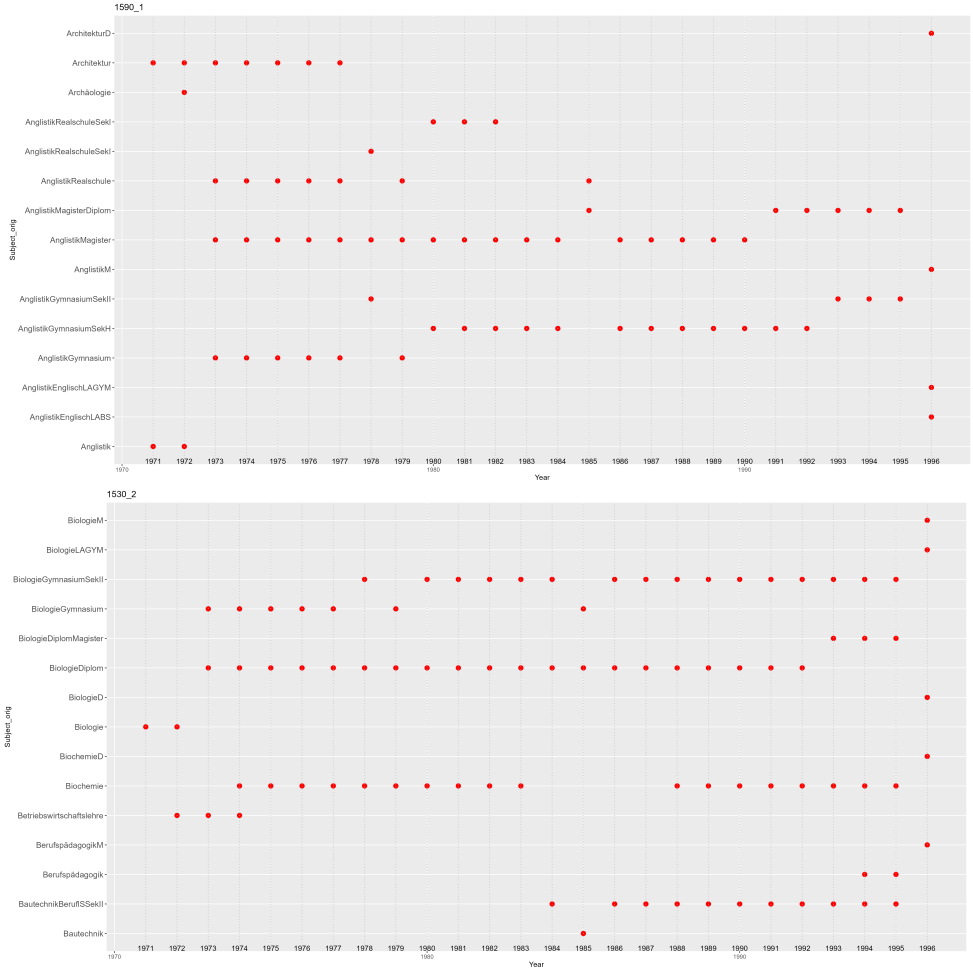
Figure 8 provides an example of **visual inspection**. It illustrates the table grid of a particular table, highlighting in blue the horizontal lines detected by the algorithm. Similar visualizations were also utilized for vertical line detection. This visualization method facilitated rapid error analysis and was employed for every table recognition instance. In cases where lines were not detected or an excessive number of lines were identified, algorithm parameters such as the minimum length of a line or the minimum distance between two lines were adjusted accordingly. The flexibility to adjust these parameters was crucial because the display format of tables was not entirely consistent over the years.

Figure 8: Visual example of grid recognition



Furthermore, the subjects offered by each university were plotted over the years, and all resulting images underwent manual review. If any data exhibited implausible fluctuations, the original and generated datasets were manually cross-checked, and the extraction algorithm was adjusted accordingly. An example of such a plot is shown in Figure 9.

Figure 9: Appearance of subjects over time



Note: The figure illustrates the temporal evolution of a subject within a particular institution, based on the name of the subject in the guides. The upper panel exemplifies data from the University of Stuttgart, while the lower panel presents data from TU Darmstadt. Notably, variations or discontinuities in the naming of subjects may occur over time.

A manual comparison was conducted between the extractions and the original sources to ensure data accuracy. These comparisons encompassed not only the axis labels but also an examination of the entire tables. As the **axis labels** were recorded independently from the table grid recognition process, a comparison was made between the number of manually recorded axis labels and the number of recognized rows or columns. If discrepancies arose where the number of manually recorded axis labels did not align with the number of recognized rows or columns, both sides were thoroughly examined. Depending on the source of the error, adjustments were made either to the axis labels or to the table grid recognition for the affected tables. This iterative approach ensured alignment between the axis labels and the table grid, thus maintaining data accuracy and integrity.

As an additional measure to assess the quality of the data, the resulting tables underwent **manual comparison with the original tables** in the guides. During each evaluation, two to six rows and two to six columns of each table were randomly selected and cross-checked against the corresponding rows and columns on different pages of the original tables. On average, these randomly selected cells constituted 5.8% of all cells. To assess whether random selection of cells is an appropriate method for evaluating the quality of the data extraction algorithm, the table for the year 1992 was chosen due to its lower scan quality, and a larger number of cells were inspected. Despite the increased number of checked cells, the

error rates remained within the previously stated range, even considering the lower quality of the guide scans. This suggests that random selection of cells is sufficient for determining error rates and allows for extrapolation of error rates across the entire dataset.

2.3 Additional data preparation

Further data enrichment steps were required to create the final version of RWI-UNI-SUBJECTS. These steps involved harmonizing the names of higher education institutions and subjects. Additional columns were furthermore added to map all subjects to the German academic coding system provided by Destatis (2023) as well as to the international subject classifications provided by UNESCO (ISCED-F, 2013). Finally, locations and municipality codes were added to the data.

Names and identifiers of higher education institutions

In the first step, all axis labels of universities and UAS were manually adjusted to harmonize any reading errors, duplicates, and inconsistent names over time. In a second step, higher education institution codes from the German Federal Statistical Office were matched to the data as they provide the unique identifiers used in German higher education statistics (Destatis, 2022).

In Destatis (2022), the Federal Statistical Office provides names and unique identifiers for all German higher education institutions. In total, 413 different names from the guides had to be matched to the identifiers from Destatis (2022). In roughly 70% of cases, an assignment was easy to implement due to the high similarity in names between the guides and Destatis (2022). For instance, "Düsseldorf U" could be assigned to "University of Düsseldorf" without further ado. For the remaining institutions, further investigation was required based on their appearance and the subjects offered. For example, "Berlin Sozial FH", could be correctly assigned to "Alice-Salomon-Hochschule Berlin (FH)" based on the fields offered. In a few exceptional cases, the guides were manually checked for additional information regarding the relevant higher education institutions. For example, each guide includes contact details for each university's disability officer, allowing for the recognition of the abbreviated name of the university. This meticulous approach ensured accurate identification and assignment of institutions where automated methods were insufficient.

During the process of adding unique identifiers based on Destatis (2022) to the dataset, several special cases occurred:

1. **Change of names:** Name changes may occur. In such cases, the current name and, therefore, the current identifier are assigned, and the variable `hei_change` is set to "Change of institution name". Additionally, the previous name is included in the dataset.

2. **Mergers and integration:** A merger of two institutions or integration of one institution into another may occur. In such cases, the current name is assigned, and if the separate location still maintains its own identifier, that identifier is utilized. The variable `hei_change` is set to "Merger or integration (location retains HE number)" in this scenario. If an integration takes place and the separate campus no longer retains its own identifier, the current name is assigned, and variable `hei_change` is set to "Merger or integration (location loses HE number)". In both instances, the last previously valid name is separately included.

3. **Separate campus:** It may occur that a higher education institution offers a small and very limited range of subjects at one location, resulting in this campus never being assigned its own identifier. In such cases, the indentation in the guides, as depicted in Figure 10, was utilized to allocate the campus to its associated institution. For instance, the example in Figure 10 provides

that "Gummersbach" is a campus of UAS Cologne. In addition, the websites of higher education institutions were manually checked for additional information. If the location could be identified as a campus, the name and thus the identifier of the associated higher education institution are assigned, and `hei_change` is set to "Separate campus without its own HE number", indicating the existence of a campus without its own identifier.

Figure 10: Identification of a separate campus by indentation in the guides

The campus "Gummersbach" of Cologne UAS is in the second column of the table. It is just a campus, not an independent higher education institution, as indicated by the intended position. Further to the right are other campuses, such as "Höxter", "Meschede", and "Soest", all of which are part of the University of Paderborn.

4. **Comprehensive University:** Comprehensive Universities (in German: *"Gesamthochschulen (GH)"*) was a form of higher education institution that combined features of universities and UAS. These existed at a few university locations in Germany between 1971 and 2003. Because comprehensive universities combined features of universities and UAS, they appear in the guides as both a university and a UAS, each offering respective subjects. To denote this exceptional circumstance, the variable `hei_change` is set to "Former comprehensive university". Additionally, the institution is assigned the most recent name.

Locations

The names of the institutions were then used to extract the location, specifically the name of the city, which is often part of the institution's name. Subsequently, the official municipality code (in German: "Amtlicher Gemeindeschlüssel, AGS") in accordance with GeoBasis-DE / BKG 2013 (as of the territorial status of 31 December 2013) was added in an additional column. The municipality code enables, for instance, direct linkage to all data in the National Education Panel (NEPS), given the identical territorial status in both datasets.

Subjects and the academic coding system

Similarly, the names of the subjects were manually harmonized to eliminate reading errors and spelling inconsistencies. Then, the official codes from the German academic coding system were appended to the subject names as provided by Destatis (2023).

In Germany, the Federal Statistical Office provides an academic coding system (in German: *"Fächersystematik"*) that distinguishes 276 subjects, summarised into 64 subject areas and further aggregated into 9 subject groups. The complete assignment of subject names from the study guides to those of the academic coding system, is shown in Appendix A. Additionally, we assign an international classification to each subject using the *"International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F)"*, developed by UNESCO in 2013. The classification by ISCED-F allows for

the comparison of educational fields across countries and systems. For more information, see ISCED-F (2013).

Because there is no official mapping between German Destatis subject codes and the international ISCED-F codes, a few subjects were difficult to match. These cases mostly related to interdisciplinary programs: In fact, the same universities may assign the same subject to different ISCED-F codes. Figure 11 illustrates such a case by providing a screenshot of the ISCED-F classification used by TU Darmstadt to categorize its offered subjects. Here, Industrial Engineering (in German: *Wirtschaftsingenieurwesen*) may be classified as ISCED-F code 041 (Business and Administration) if the curriculum focuses primarily on business or economics, or as code 071 (Engineering and Engineering Trades) if technical courses predominate (Technical University Darmstadt, 2025). In such cases, we selected the classification that most closely represents the program.

(FB01) Wirtschaftsingenieurwesen	041 - Business and administration (bei überwiegender Belegung wirtschaftswissen-schaftlicher Kurse) **oder** 071 - Engineering and engineering trades (bei überwiegender Belegung technischer Kurse)

Figure 11: Screenshot of the ISCED-F by TU Darmstadt

Note: Own screenshot from Technical University Darmstadt (2025) on July 16, 2025.

English Translation

All variables were translated into English to improve the dataset's usability for an international audience. For the translation process of the subjects, we used the official Destatis (2020) English translation list to convert all German terms related to subjects, subject areas, and subject groups into English. However, since some subject matches were ambiguous, the `Destatis_subject field` occasionally contained missing values. To address these cases, we additionally translated all original subject terms from the study guides (subject) using the pre-trained MarianMT language model Junczys-Dowmunt et al. (2018) (*Helsinki-NLP/opus-mt-de-en* by Tiedemann and Thottingal (2020)). This model was further fine-tuned using the official Destatis translations to ensure that machine-translated subject titles closely matched Destatis terminology.

2.4 Data quality and limitations

Data quality

At the end of the intensive data preparation exercises, a final check was conducted to assess the quality of the dataset. Similar to the interim checks conducted, rows and columns of the final dataset were randomly chosen and compared to the original guides. The resulting error rate was determined as the proportion of incorrect entries overall randomly chosen cells in a final manual comparison. The extrapolated average error rate of the final dataset is 0.2%, meaning that one in every 500 entries may contain an error. Please refer to Table 3 for more details.

Limitations

Structural breaks within the guides introduce complexities to the comparative analysis of subject offerings over time. These breaks encompass significant events such as German reunification, the aggregation of UAS at the city level until 1985, the introduction of tables dedicated to highly specific higher education institutions, and the reorganized presentation of teacher training programs.

- **German reunification:** The German reunification united Germany and the German Democratic Republic (GDR) in 1990. Data on higher education institutions in the former GDR is not available (see also Figure 3).

- **Display of UAS before 1985:** Until 1985, the guides only provided information on UAS aggregated at the city level. In cases where two or more UAS were situated within the same city, the guides consolidated the information at the city level, without distinguishing between individual UAS. From 1986 onwards, it is possible to distinguish two or more UAS located in the same city

- **Introduction of new tables:** The two main tables from the study guides are those for universities and UAS. The current dataset contains the universe of study programs displayed within these main tables. However, from 1978 onwards, additional, much smaller tables were introduced. These additional tables contain the study programs of church colleges, colleges and academies of fine arts and other highly specific types of institutions. The introduction of relevant additional tables was as follows:

  - 1978: *Church colleges*, *theological colleges*, *colleges and academies of fine arts*, *colleges of music*, and *other colleges (e.g., sports college, academy of film and television, college of political science)* were separately incorporated. They are not part of the current dataset.
  - 1986: *UAS for public administration* were separately incorporated. They are not part of the current dataset.
  - 1991: *Universities of the German armed forces* were introduced as a separate table in the study guides. They are not part of the current dataset.
  - 1995: A table detailing subjects available through distance learning was introduced. It is not part of the current dataset.
  - 1996: Information on the *universities of the German armed forces, church colleges, philosophical-theological colleges, and those listed in the table of other higher education institutions* were incorporated into the main table of universities. This means that they were assigned the type "university" and are part of the current dataset, which explains the significant increase observed in the latter in 1996, as depicted in Figure 3.

  The study programs from the separate tables will soon be made available as supplementary material to the RWI-UNI-SUBJECTS dataset. This extension will follow the same format but will be approximately 5% of the size of the current dataset. However, it will not include data on distance learning or UAS for public administration, due to their specific characteristics. The table on UAS for public administration mainly features subjects that are only loosely connected to conventional study programs. The distance learning table can not be mapped on a regional level, making it incompatible with the rest of the dataset.

- **University of Kassel in 1972:** In the guide of 1972, the University of Kassel is missing. Most likely, it has been inadvertently omitted. However, information regarding available subjects for this university is included in the guides for the years 1971 and 1973, and all subsequent years after 1973.

Missings in Variables

- **Institutions:** In the variables `Destatis_hei_name` and `Destatis_hei_number`, 9.90% of the entries are missing. Nearly all of these (99.96%) are attributable to the fact that, prior to 1985, universities of applied sciences were listed in the study guides only by location, making it impossible to

identify the exact institution responsible for each program. In these cases, the entries are marked with the suffix "total" in the `hei_name` variable. The remaining 0.04% of missing values stem from two institutions whose names could not be definitively linked to any official Destatis code or name, although location and institution names from the study guides are available.

- **Subjects:** All observations have been successfully assigned to the broadest Destatis category (`Destatis_subject_group`). However, at more detailed levels, some observations remain unmatched: 1.49% at the `Destatis_subject` level and 0.33% at the `Destatis_subject_area` level. Examples of these missings include museum studies and crystallography.

3 References

BKG. (2023). *Shapefiles Deutschland.*

Bock, K. H. (1973). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Bundesanstalt für Arbeit, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1974). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Bundesanstalt für Arbeit, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1975). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Bundesanstalt für Arbeit, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1976). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1977). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1978). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1979). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1980). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1981). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1982). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1983). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1984). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1985). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1986). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1987). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1988). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1989). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1990). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1991). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1992). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1993). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1994). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1995). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bock, K. H. (1996). *Studien- und Berufswahl - Entscheidungshilfen für Abiturienten und Absolventen der Fachoberschulen.* (B.-L.-K. für Bildungsplanung und Forschungsförderung und Bundesanstalt für Arbeite, Ed.) (No. 1). Verlag Karl Heinrich Bock.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Destatis. (2020). *Systematik der Fächergruppen, Studienbereiche und Studienfächer in Englisch [Classification of subject areas, subject groups and subjects in English].* (Available upon request, Stand: 2020/2021)

Destatis. (2022). *Schlüsselverzeichnisse für die Studenten-, Prüfungsstatistik, Promovierendenstatistik. Stand: 2021/2022 & SS 2022.*

Destatis. (2023). *Systematik der Fächergruppen, Studienbereiche und Studienfächer.*

German Centre for Higher Education Research and Science Studies (DZHW). (2020). *DZHW Graduate Survey Series.* Study Series: DZHW Graduate Survey Series. (https://doi.org/10.21249/DZHW:gra1989:2.0.0)

Hirschfeld, G. (1971). *Studium und Beruf - Informationen für Abiturienten und Absolventen der Fachoberschulen* (B.-L.-K. für Bildungsplanung und Bundesanstalt für Arbeit, Ed.) (No. 1). aspekte verlag gmbh.

Hirschfeld, G. (1972). *Studium und Beruf - Informationen für Abiturienten und Absolventen der Fachoberschulen* (B.-L.-K. für Bildungsplanung und Bundesanstalt für Arbeit, Ed.) (No. 1). aspekte verlag gmbh.

ISCED-F. (2013). *International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F 2013)– Detailed Field Descriptions* (Tech. Rep.). Montréal: UNESCO Institute for Statistics. (https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-fields-of-education-and-training-2013-detailed-field-descriptions-2015-en.pdf)

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., ... others (2018). Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Meisel, H. (1978). *Die Deutsche Berufsberatung – Gesamtüberblick.* Stuttgart, Berlin, Köln, Mainz: Kohlhammer.

NEPS-Netzwerk. (2023). *Nationales Bildungspanel, Scientific Use File der Startkohorte Erwachsene.* Bamberg. (https://doi.org/10.5157/NEPS:SC6:14.0.0)

Schmucker, A., Seth, S., & vom Berge, P. (2023). *Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) 1975 - 2021* (Tech. Rep. No. 02/2023). Nürnberg: FDZ Datenreport. (https://doi.org/10.5164/IAB.FDZD.2302.de.v1, language = de)

Technical University Darmstadt. (2025). *International Standard Classification of Education (ISCED) – Zuordnung der ISCED-Codes zu den Studienfächern* (Technical Report). Technische Universität Darmstadt. (https://www.intern.tu-darmstadt.de/media/dezernat_viii_medien/viii_c_/ig_internationales/studiereninternational/dokumenteinout/downloads_1/ISCEDZuordnung.pdf, Accessed on July 16, 2025.)

Tiedemann, J., & Thottingal, S. (2020). Opus-mt–building open translation services for the world. In *Annual conference of the european association for machine translation* (pp. 479–480).

University of Bologna. (2024). *Degree Programmes Data – University of Bologna.* Retrieved from https://dati.unibo.it/dataset/degree-programmes (Catalogue of First and Second Cycle Degree Programmes of the University of Bologna)

U.S. Department of Education. (2024). *College Scorecard Data.* Retrieved from https://collegescorecard.ed.gov/data/ (Data on student enrollment, completion, debt, repayment, and earnings for U.S. institutions)

A Academic Coding System

The guides to "Study and Career Choice" from 1971 to 1996 differentiate between subjects. These are assigned to subjects, subject areas, and subject groups based on the academic coding system from Destatis (2023). Notably, no subject was assigned to the ninth subject group from the academic coding system, labeled "Outside the subject area classification". Below is an example of the assignment for each

subject group. The hierarchical structure delineates the subject group, subject area, and corresponding subjects from the academic coding system, along with their respective assigned subject names from the guides.[3]

1 Geisteswissenschaften [Humanities]

(a) 02 Evang. Theologie, -Religionslehre [Protestant Theology, Protes. Religious Education]

i. 053 Evang. Theologie, -Religionslehre [Protestant Theology, Protes. Religious Education]

Name as in the guide: *Theologie ev.* [Protes. Theology]

...

2 Sport [Sports]

(a) 22 Sport, Sportwissenschaft [Sports, Sports Science]

i. 029 Sportwissenschaft [Sports Science]:

Names as in the guide: *Sport* [Sports], *Sportwissenschaft* [Sports Science]

...

3 Rechts-, Wirtschafts- und Sozialwissenschaften [Law, Economics and Social Sciences]

(a) 25 Politikwissenschaft [Political Science]

i. 129 Politikwissenschaft/Politologie [Political Science]:

Names as in the guide: *Politik/Sozialkunde* [Political Science, Social Studies], *Politikwissenschaft* [Political Science], *Politikwissenschaft/Politologie* [Political Science], *Politologie* [Political Science]

...

4 Mathematik, Naturwissenschaften [Mathematics, Natural Sciences]

(a) 37 Mathematik [Mathematics]

i. 105 Mathematik [Mathematics]:

Names as in the guide: *Mathematik* [Mathematics], *Mathematik (Diplom)* [Mathematics Diploma], *Mathematik (Diplom/Magister)* [Mathematics (Diploma/Magister], *Mathematik (Gymnasium)* [Mathematics (Gymnasium)], *Mathematik (Realschule) [Mathematics (Realschule)]*

...

5 Humanmedizin/Gesundheitswissenschaften [Human Medicine/Health Sciences]

(a) 50 Zahnmedizin [Dentistry]

i. 185 Zahnmedizin [Dentistry]:

Names as in the guide: *Zahnmedizin* [Dentistry], *Zahnmedizin (Stomatologie)* [Dentistry (Stomatology)]

...

[3]The German word *Gymnasium* refers to the academic track in teacher education, while *Realschule* denotes the non-academic track. As Destatis (2020) does not provide official English translations for these terms, we use the original German words. The same is true for the other school tracks.

7 Agrar-, Forst- und Ernährungswissenschaften, Veterinärmedizin [Agricultural, Forest and Nutritional Sciences, Veterinary Medicine]

(a) 58 Agrarwissenschaften, Lebensmittel- und Getränketechnologie [Agricultural Science, Food and Beverage Technology]

i. 003 Agrarwissenschaft/Landwirtschaft [Agricultural Science/Agriculture]:

Names as in the guide: *Agrarwissenschaft* [Agricultural Science], *Landwirtschaft* [Agriculture], *Landwirtschaft/Agrarwirtschaft* [Agriculture/Agricultural Science]

...

8 Ingenieurwissenschaften [Engineering Sciences]

(a) 63 Machinenbau/Verfahrenstechnik [Mechanical Engineering/Process Engineering]

i. 104 Maschinenbau/-wesen [Mechanical Engineering]:

Names as in the guide: *Maschinenbau* [Mechanical Engineering], *Maschinenbau/Stahlbau* [Mechanical Engineering/Steel Construction], *Maschinenbau allgemein* [Mechanical Engineering (general)], *Kraftfahrzeugbau* [Automotive Engineering]

...

9 Kunst, Kunstwissenschaft [Art, Art Theory]

(a) 74 Kunst, Kunstwissenschaften allgemein [Art, Art Theory (general)]

i. 091 Kunsterziehung [Art Education]:

Names as in the guide: *Kunst/Kunsterziehung* [Art, Art Education], *Kunst/freie Kunstpädagogik* [Art, Free Art Education], *Kunst/-pädagogik* [Art Education], *Kunst (Gymnasium)* [Art (Gymnasium)]

...

B Details on temporal evolution of institutions

B.1 Number of subjects and institutions over time

B.1.1 Number of subjects over time

Table 1: Number of subjects by year and type of higher education institution

Year	UAS	Uni	UAS & Uni
1971	586	1349	1935
1972	476	1202	1678
1973	666	2462	3128
1974	692	2442	3134
1975	696	2582	3278
1976	691	2580	3271
1977	725	2699	3424
1978	697	2886	3583
1979	711	2938	3649
1980	702	2991	3693
1981	695	3010	3705
1982	754	3140	3894
1983	763	3060	3823
1984	776	3107	3883
1985	788	3194	3982
1986	831	3155	3986
1987	840	3184	4024
1988	882	3217	4099
1989	905	3216	4121
1990	958	3335	4293
1991	1059	3785	4844
1992	1175	4086	5261
1993	1327	4270	5597
1994	1370	4389	5759
1995	1419	4372	5791
1996	1510	5962	7472

Note: The table summarizes the number of subjects by year and type of institution.

B.1.2 Number of institutions over time

Table 2: Number of higher education institutions by institutional type over time

Year	Number of Unis	Number of UAS	Number of municipalities with UAS
1971	44	-	95
1972	44	-	102
1973	49	-	100
1974	52	-	103
1975	52	-	104
1976	55	-	104
1977	55	-	104
1978	56	-	106
1979	57	-	109
1980	56	-	107
1981	58	-	105
1982	59	-	105
1983	61	-	105
1984	61	-	106
1985	61	-	105
1986	61	135	106
1987	61	136	106
1988	61	137	106
1989	61	135	107
1990	61	136	107
1991	74	152	121
1992	80	169	132
1993	79	177	137
1994	79	184	143
1995	79	193	151
1996	103	198	155

Note: The table shows the number of higher education institutions based on the count of unique higher education institutions. Since UAS prior to 1986 are only listed by place of study, the number of municipalities with UAS is presented separately.

C Error rates per year

Table 3 provides the number of randomly selected cells, the number of incorrect entries, and the resulting error rates for each guide. Depending on the quality of the tables in the guides as well as on the quality of scans, the error rate varies over the years and ranges from 0% to 1.27%. The extrapolated average error rate is 0.2%, meaning that one in every 500 entries may contain an error.

Table 3: Error rate of checked cells per year and institution

Year	Universities			UAS		
	Number of checked cells	Number of incorrect entries	Error rate	Number of checked cells	Number of incorrect entries	Error rate
1971	396	3	0.75%	390	0	0.00%
1972	299	1	0.33%	395	0	0.00%
1973	392	5	1.27%	400	0	0.00%
1974	591	7	1.18%	398	2	0.50%
1975	597	1	0.16%	399	3	0.75%
1976	612	0	0.00%	421	1	0.24%
1977	410	3	0.73%	528	0	0.00%
1978	678	1	0.14%	561	0	0.00%
1979	684	0	0.00%	564	0	0.00%
1980	513	2	0.38%	540	1	0.19%
1981	690	4	0.58%	429	1	0.23%
1982	582	2	0.34%	444	1	0.23%
1983	352	0	0.00%	549	1	0.18%
1984	470	0	0.00%	444	1	0.23%
1985	474	0	0.00%	442	0	0.00%
1986	530	3	0.57%	512	0	0.00%
1987	590	0	0.00%	588	0	0.00%
1988	594	3	0.51%	669	0	0.00%
1989	536	0	0.00%	672	0	0.00%
1990	735	0	0.00%	558	0	0.00%
1991	558	0	0.00%	864	0	0.00%
1992	1973	17	0.86%	634	0	0.00%
1993	738	0	0.00%	646	0	0.00%
1994	819	1	0.12%	662	0	0.00%
1995	548	0	0.00%	873	3	0.34%
1996	903	0	0.00%	598	0	0.00%

Note: The table provides a summary of the results from the random manual review of item identification for UAS and universities. The average error rate across all years from 1971 to 1996 is 0.20%. Notably, the guide from 1992 exhibited low scan quality, necessitating a broader selection of cells for quality control.

D Codebook

Variable name	**Description**
Variables extracted from the study guides:	
`year`	Publication year of the study guide and thereby observation (1971 until 1996)

Variable name	Description
type	Institutional type, e.g., *university* or *university of applied sciences*
hei_name	Institution's name as provided by the guide
subject	Original study program name as provided by the guide
study_type	Variable indicating the type of the study modes: *Full study* *Full study Winter term (WS) required* *Full study Winter term (WS) recommended* *Full study Summer term (SS) required* *Full study admission-restricted* *Full study admission-restricted Winter term (WS) required* *Full study admission-restricted Winter term (WS) recommended* *Specialization* *Specialization Winter term (WS) required* *Specialization Winter term (WS) recommended* *Specialization admission-restricted* *Advanced study* *Advanced study Winter term (WS) required* *Advanced study Winter term (WS) recommended* *Advanced study admission-restricted* *Partial study* *Partial study Winter term (WS) required* *Partial study Winter term (WS) recommended* *Partial study starting from* *Partial study until* *Minor subject* *Minor subject Winter term (WS) required* *Minor subject Winter term (WS) recommended* *No new students* *No new students soon*
Enrichment to ensure effective use:	
Destatis_hei_number	Institution's code based on Destatis (2022)
Destatis_hei_name	Institution's name based on Destatis (2022)
Destatis_hei_name_last	Last previous institution's name based on Destatis (2022) in case of observed name changes
exact_hei_name	Binary variable equal to 1 if the institution's name, as provided by the guide, refers to an exact institution, and 0 if it refers to a location-level aggregation (all universities of applied sciences prior to 1986)

Variable name	Description
hei_change	Categorical Variable indicating changes in the institution characteristics and Destatis_hei_number: *No observed change* *Change of institution name* *Merger or integration (location retains HE number)* *Merger or integration (location loses HE number)* *Separate campus without its own HE number* *Former comprehensive university*
Destatis_subject	Study program based on Destatis (2023)
Destatis_subject_area	Subject area based on Destatis (2023), e.g., engineering
Destatis_subject_group	Subject group based on Destatis (2023), e.g., computer sciences
Destatis_subject_code	Subject code based on Destatis (2023), e.g., business informatics
Destatis_subject_area_code	Subject area code based on Destatis (2023)
Destatis_subject_group_code	Subject group code based on Destatis (2023)
location_name	Location (city) of the institution
BKG_municipality_code	Municipality code as of December 31, 2013 (BKG, 2023)
ISCED_detailed_field	Detailed field classification based on ISCED-F (2013) , e.g., food processing
ISCED_narrow_field	Narrow field classification based on ISCED-F (2013) , e.g., manufacturing and processing
ISCED_broad_field	Broad field classification based on ISCED-F (2013), e.g., engineering, manufacturing and construction
ISCED_detailed_field_code	Detailed field code based on ISCED-F (2013)
ISCED_narrow_field_code	Narrow field code based on ISCED-F (2013)
ISCED_broad_field_code	Broad field code based on ISCED-F (2013)
subject_EN	English translation of the original subject
Destatis_subject_EN	English translation of the subject based on Destatis (2020)
Destatis_subject_area_EN	English translation of the subject area based on Destatis (2020)
Destatis_subject_group_EN	English translation of the subject group based on Destatis (2020)
author	The individual or group responsible for writing the study guide
commissioning_body	The organization responsible for overseeing or commissioning the study guide. In this dataset, this is exclusively the *"Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK)"*, a joint commission of the German federal and state governments for educational planning and research funding.
title	The full title of the study guide as it appears on the publication
publisher	The publishing institution responsible for the production and distribution of the study guide